EMANUEL CZUBER

DIE STATISTISCHEN FORSCHUNGSMETHODEN

DRITTE ERWEITERTE AUFLAGE

HERAUSGEGEBEN VON

F. BURKHARDT

PROFESSOR AN DER UNIVERSITÄT LEIPZIG

MIT 38 FIGUREN IM TEXT

19 38

VERLAG VON L. W. SEIDEL & SOHN IN WIEN

Aus dem Vorwort zur ersten Auflage.

Zur Abfassung dieses Buches bin ich hauptsächlich durch das Erscheinen von G. Udny Yule's „An Introduction to the Theory of Statistics“[1]) veranlaßt worden. Das Buch Yūle's ist 1911 zum erstenmal erschienen und hat seither vier weitere Auflagen erlebt; dies allein spricht für das Interesse und den Eifer, mit welchem statistische Studien betrieben werden. Einen Hinweis darauf hat F. Klein in seiner „Festrede zum 20. Stiftungstage der Göttinger Vereinigung zur Förderung der Angewandten Physik und Mathematik (1918)[2])“ gemacht.

Dem praktischen Ziele entsprechend, habe ich in der vorliegenden Schrift das Hauptgewicht auf reichliche Anwendungen gelegt. Den dazu nötigen Stoff entnahm ich außer Yule selbst verschiedenen anderen Quellen und war dabei darauf bedacht, daß möglichst viele der Gebiete vertreten seien, die sich heute schon der mathematischen Methoden bedienen. In rechnerischer Beziehung ist das Summenverfahren vor anderen Rechnungsweisen bevorzugt worden; doch kamen auch andere Arten der Rechnungsanlage zur Geltung.

Der nicht immer einwandfreien Übertragung der Fehlertheorie auf statistische Untersuchungen mußten einige Worte gewidmet werden; insbesondere ist bei der mathematischen Bearbeitung land- und forstwirtschaftlicher Versuche von der auf Wahrscheinlichkeitsrechnung gegründeten Fehlertheorie ein übermäßiger, oft ungerechtfertigter Gebrauch gemacht worden.

Entsprechend dem weiten Kreise, an den sich das Buch wendet, sind die theoretischen Erörterungen so ausführlich gehalten, daß ihnen leicht gefolgt werden kann.

Gnigl im Salzburgischen, 4. April 1920.

E. Czuber.

[1]) London, Charles Griffin and Company.

[2]) Jahresbericht der Deutschen Mathematiker-Vereinigung, 27. Band, S. 217, insbesondere 223.

Vorwort zur dritten Auflage.

Bei der Bearbeitung der dritten Auflage des vorliegenden Buches habe ich die Gegenstände, auf die Czuber die statistischen Forschungsmethoden angewandt hat, im wesentlichen beibehalten. Die Zahlenbeispiele zu diesen Gegenständen ersetzte ich durch neues Zahlenmaterial, und zwar so weit wie möglich aus der amtlichen Statistik. Bei einzelnen Gegenständen jedoch standen mir die entsprechenden neuen Zahlen trotz eifrigster Bemühungen nicht zur Verfügung, so daß ich das von Czuber verwandte Zahlenmaterial, das sich auch in einigen neueren statistischen Lehrbüchern und in neueren statistischen Aufsätzen vorfindet, übernehmen mußte.

Bei der Behandlung der beibehaltenen Gegenstände ist im allgemeinen an der von Czuber gegebenen textlichen und mathematischen Darstellung nur wenig geändert worden. Im besonderen ist an der mitunter sehr ausführlichen, an schwierigeren Punkten sogar sehr breiten Darstellungsweise nur wenig gekürzt worden, um das Lesen dieses Buches nicht zu erschweren.

Neu aufgenommen wurden folgende Gegenstände: Bevölkerungsschwerpunkt, mittlere Bevölkerungszahl, Verhältniszahlen, schärfere Methoden zur Berechnung von Beziehungszahlen, Sterblichkeit im ersten Lebensjahre mit Unterscheidung der Legitimität, Vergleichung von Abgangswahrscheinlichkeiten, Bestimmung des Trends, Trendlinien in der bevölkerungsstatistischen Forschung, Bestimmung der Saisonschwankungen, Finanzausgleich. Bei der Auswahl dieser neu aufgenommenen Gegenstände war der Gesichtspunkt maßgebend, möglichst die zu berücksichtigen, die in der statistischen Praxis von Wichtigkeit sind. Dabei mußten die sachlichen Ausführungen über diese Gegenstände mit Rücksicht auf den zur Verfügung stehenden Raum ziemlich kurz gehalten werden. Um den Leser zu weiteren Forschungen anzuregen, sind außerdem Hinweise auf solche Fragen gegeben worden, zu deren Lösung mathematisch-statistische Methoden neuerdings mit Erfolg angewandt werden.

Bei der Darstellung der verschiedenen Methoden wurde immer darauf Wert gelegt, sowohl die mathematische als auch die logische Seite der statistischen Forschungsweise herauszuarbeiten.

Nicht behandelt wurde die Anwendung der statistischen Forschungsmethoden in der theoretischen Physik, da die Gesamtheiten, auf die man hier statistische Methoden ansetzt, nicht durch Auszählen gewonnen werden.

Infolge der Vielgestaltigkeit der Anwendungsgebiete der statistischen Methoden war es in einzelnen Fällen nicht zu umgehen, bereits verwandte Buchstaben bei nachfolgenden Betrachtungen in anderem Sinne zu benutzen.

Bei besonders schwierigen Betrachtungen werden an einigen Stellen zunächst Beispiele gebracht und die diesen Beispielen zugrunde liegenden allgemeinen Methoden später entwickelt. Die erforderlichen Hinweise sind angebracht worden.

In einzelnen Beispielen ist die Berechnung von Dezimalstellen etwas weit getrieben. Dabei war der Gesichtspunkt maßgebend, das methodisch Bemerkenswerte auch in den Zahlenbeispielen deutlich zu zeigen. Um den Fortgang eines mathematischen Ausdrucks, der sich über zwei Zeilen erstreckt, deutlich zu machen, wurde das Vorzeichen am Anfang der zweiten Zeile wiederholt. Zur Vereinfachung der Schreibweise solcher Ausdrücke ist bei Briggischen Logarithmen an Stelle von log später lg geschrieben worden.

Bei der Bearbeitung dieser Auflage ließ ich mich von dem Gedanken leiten zu zeigen, wie die mathematisch-statistischen Methoden auf den verschiedenen Lebensgebieten fruchtbringend angewandt werden können. Voraussetzung ist hierbei: sorgfältige Erhebung, genaue Prüfung, exakte Aufbereitung des statistischen Materials.

Besonders danke ich Fräulein Studienassessor J. Barthel für wissenschaftliche Mitarbeit und tatkräftige Unterstützung.

Der Verlagsbuchhandlung bin ich für ihr bereitwilliges Eingehen auf alle meine Wünsche zu Dank verpflichtet.

Leipzig, im Januar 1938.

F. Burkhardt.

INHALTSVERZEICHNIS.

Zweiter Abschnitt.

Theorie der veränderlichen Merkmale.

§ 1. Die Verteilungen in Kollektiven.

§ 2. Mittelwerte.

§ 3. Verhältniszahlen.

§ 4. Streuungsmaße.

§ 5. Korrelation zwischen zwei Variablen.

Theorie.

§ 6. Korrelation zwischen zwei Variablen.

Praktische Durchführung.

§ 2. **Die binomiale Verteilung.**

§ 3. **Die normale Häufigkeitskurve.**

§ 4. **Normale Korrelation.**

In den Beispielen behandelte Stoffe.

(Die Nummern beziehen sich auf die Artikel.)

Einleitung.

1. Im Gegensatze zu solchen Forschungsweisen, welche das innere Wesen der Dinge betreffen, wie die Methoden physikalischer, chemischer, physiologischer Forschung, betreffen die statistischen Methoden bloß das Äußere der Dinge, und es ist wesentlich für sie, daß sich ihre Ergebnisse in Zahl und Maß ausdrücken, also quantitativer Natur sind.

Im ersten Falle richtet sich die Frage darauf, wie viele der untersuchten Gegenstände einer Gesamtheit ein bestimmtes Merkmal oder eine bestimmte Merkmalgruppe aufweisen, und das erfordert ein Zählen. Die Merkmale sind dabei qualitativer Natur oder haben als qualitativ zu gelten, selbst wenn sie sich auf Quantitatives beziehen. So bezeichnen Farben, wenn man nicht auf fernliegende physikalische Betrachtungen zurückgreifen will, etwas Qualitatives. Groß und klein, schwer und leicht, dick und dünn und ähnliches beziehen sich wohl auf Quantitatives, können aber ohne Vornahme einer weiter zu verwendenden Messung wie qualitative Eigenschaften behandelt werden.

Die Frage, die z. B. gegenüber einem Gegenstande gestellt wird, lautet das eine Mal: Ist er so oder so? Das andere Mal fragt man: Wie groß, wie schwer ist der Gegenstand? Das erste Mal kommt das Quantitative erst durch das Zählen herein, das zweite Mal ist es schon vor einer vorzunehmenden Zählung da und wird durch die Messung oder Wägung herbeigeführt.

Man hat also zwei Arten von Erhebungen zu unterscheiden: Bei der ersten Art handelt es sich um die Feststellung des Vorhandenseins oder Fehlens eines bestimmten Merkmals oder mehrerer solcher, um die Zusammenfassung zu Klassen und um deren Größenbestimmung durch Zählung; bei der anderen Art handelt es sich um die zahlenmäßige Bestimmung des Grades eines variablen Merkmals, das entweder von Anfang an quantitativen Charakter besitzt oder auf künstlichem Wege auf einen solchen zurückgeführt worden ist.

Wenn z. B. bei einer Volkszählung die Personen nach Geschlecht, Familienstand, Religionszugehörigkeit, Staatsangehörigkeit und ähnlichem unterschieden werden, dann fällt dies unter die erste Art der Erhebungen. Soll der Altersaufbau einer Bevölkerung erforscht werden, dann hat man es mit Erhebungen der zweiten Art zu tun.

2. Die Gegenstände, von welchen hier die Rede ist, sind stets von gleicher Art. Doch kann die Gleichartigkeit[1]) verschiedenen Grades sein, je nachdem die Übereinstimmung in einer geringeren oder größeren Zahl von Merkmalen gefordert

[1]) Das statistische Problem der Gleichartigkeit ist eingehend von Žižek und Flaskämper untersucht worden. (Vgl. F. Žižek, Gleichartigkeit, Homogenität und Gleichwertigkeit in der Statistik. Allgemeines Statistisches Archiv, 18. Bd., 1928, S. 393 u. f., und P. Flaskämper, Das Problem der „Gleichartigkeit" in der Statistik. Allgemeines Statistisches Archiv, 19. Bd., 1929, S. 205 u. f.)

wird. Das richtet sich nach dem Zwecke der Untersuchung: der Grad der Gleichförmigkeit muß der Fragestellung angepaßt sein, soll die Antwort einen Wert haben. Zu geringe Gleichartigkeit kann dem Ergebnis jede wissenschaftliche Bedeutung nehmen.

Wenn es sich z. B. um die menschliche Körpergröße handelt, so hätte es mit Rücksicht auf die ungleiche Körperentwicklung in verschiedenen Altern und bei den beiden Geschlechtern kaum einen Sinn, wenn man ein Gemisch jugendlicher und erwachsener, männlicher und weiblicher Personen zur Grundlage nähme; mit dem gefundenen Resultate wäre kaum etwas anzufangen. Man wird, um ein wissenschaftlich brauchbares Resultat zu gewinnen, einen Personenkreis verwenden müssen, der in gewissen Merkmalen: Geschlecht, Alter, Abstammung u. a. einheitlich ist. Die Forderung der Gleichartigkeit kann unter Umständen sehr weit gehen; solche extreme Fälle kommen beispielsweise bei Erblichkeitsuntersuchungen vor, wo es nicht ausreicht, Pflanzen *einer* Art heranzuziehen, wo die Fragestellung vielmehr verlangt, daß die Pflanzen aus Samen *einer* Mutterpflanze hervorgegangen seien, und selbst unter diesen Samen kann noch eine Auslese nach bestimmten Kennzeichen notwendig sein.

Eine Zusammenfassung gleichartiger Gegenstände zum Zwecke einer statistischen Bearbeitung soll mit dem Namen *Kollektiv* (Sammelgegenstand) belegt werden; die Anzahl der in ihm vereinigten Einzelgegenstände, *Elemente*, *Objekte*, *Individuen*, *Exemplare*, *Glieder* oder ähnliches wird als *Umfang* des Kollektivs bezeichnet.

Es können die verschiedensten Gegenstände zu Kollektiven vereinigt werden. In erster Linie sei an Personen und an konkrete Gegenstände und unter diesen wieder an Naturobjekte: Tiere, Pflanzen und ihre Teile, Organe, gedacht. Auch Gegenstände der Wirtschaft und der Technik können der kollektiven Untersuchung zugeführt werden. Zur Bildung von Kollektiven nicht konkreter Gegenstände können Anlaß geben die verschiedensten Natur- und Kulturerscheinungen, die Ergebnisse gleichartiger (physikalischer, chemischer, physiologischer, medizinischer, biologischer, landwirtschaftlicher und ähnlicher) Versuche, die Ergebnisse wiederholter Messungen ein und derselben Erscheinung, die Preise von Waren zu verschiedenen Zeiten und an verschiedenen Orten, desgleichen die Arbeitslöhne. Eine erschöpfende Aufzählung alles dessen, was schon zur Kollektivbildung verwendet worden ist und noch verwendet werden kann, ist unmöglich. Daraus geht die überaus umfassende Natur des Kollektivbegriffs hervor.

Wenn die Resultate der statistischen Bearbeitung eines Kollektivs verständlich und wissenschaftlich verwertbar sein sollen, so muß eine *Beschreibung* des Kollektivs mitgegeben werden, aus der die einigenden Merkmale der Gegenstände zu erkennen sind. Eine Überschrift „Körpergrößen“ z. B. müßte als ungenügend erachtet werden; unter „Körpergrößen erwachsener männlicher Personen“ versteht man schon etwas Bestimmteres; noch schärfer wäre die Kennzeichnung „Körpergrößen männlicher Personen der Alter von 20 bis 30“, noch mehr eingeengt „Körpergrößen männlicher erwachsener Personen dieses oder jenes Stammes“ usw.

Infolge der umfassenden Natur des Kollektivbegriffs findet die statistische Forschungsweise Anwendung in vielen Wissenschaften, man kann sagen in allen, welche sich auf Erfahrung stützen, oder sie *kann* sie finden, wenn diese Art der Forschung immer weitere Verbreitung erlangt. Wenn man auch nur einige Hauptgebiete namhaft machen will, in welchen dies heute schon in größerem Umfange geschehen ist, so gelangt man bereits zu einer ansehnlichen Liste: Bevölkerungswissenschaft, Wirt-

schaftswissenschaften, Versicherungswissenschaft, Technik, Biologie, Chemie, Physik, Geographie, Astronomie, Medizin, Anthropologie, Vererbungslehre, experimentelle Psychologie [1]).

3. Das Wort „Statistik“ mit seinen Ableitungen und Zusammensetzungen hat seinen Ursprung in dem Worte „Staat (status)“. Die Staatenkunde, soweit sie gerichtet war auf die Mittel und Bedingungen der Existenz und der Entwicklung eines Staatswesens, hieß anfänglich Statistik. Einen dem jetzigen mehr genäherten Sinn erhielt das Wort, als man bei derartigen Darstellungen hauptsächlich auf solche Umstände Wert legte, die sich quantitativ und daher durch Zahlen ausdrücken lassen.

Diese enge Fassung hat aber in neuerer Zeit eine ungeahnte Erweiterung erfahren; man ist immer mehr zu der Einsicht gekommen, daß überall dort, wo Erscheinungen von besonders verwickelter Verursachung vorliegen, wo eine Trennung der Ursachen und ihre Einzelerforschung durch das Experiment entweder ausgeschlossen oder sehr erschwert ist, der einzige Weg zur Erkenntnis in der Sammlung von Tatsachen aus dem betreffenden Erscheinungsbereich ist. Mit der Sammlung allein aber ist es nicht geschehen; es ist auch notwendig, die Tatsachen kritisch zu ordnen, untereinander zu vergleichen und Schlüsse aus ihnen zu ziehen, gegebenenfalls unter Aufstellung von Hypothesen. Da es sich dabei im wesentlichen um ein Operieren mit Zahlen handelt, so leistet die Mathematik außerordentlich wertvolle Dienste.

In diesem erweiterten Sinne bedeutet „Statistik“ die planmäßige Sammlung und Ordnung von Massenerscheinungen zu dem Zwecke, aus ihrem zahlenmäßigen Auftreten Schlüsse zu ziehen, die zur Durchleuchtung des Erscheinungsgebietes, zur Auffindung von inneren Zusammenhängen und zum Erforschen der beherrschenden Ursachen dienen können. Die mathematischen Verfahren bilden in ihrer Gesamtheit die mathematische Statistik, die einen Teil der theoretischen Statistik darstellt. Der theoretischen Statistik steht die praktische Statistik gegenüber, die auf die Erhebung, Sammlung und Ordnung der Tatsachen gerichtet ist. Doch sind dies keine scharf getrennten und voneinander unabhängigen Betätigungen. Die praktische Statistik bedarf der Leitung durch die theoretische und diese wieder erhält von der ersteren das Material zu ihren Untersuchungen und neue Fragestellungen.

[1]) Al. A. Tschuprow spricht in dieser Hinsicht von einem Siegeszug der Statistik. (Das Gesetz der großen Zahlen und der stochastisch-statistische Standpunkt in der modernen Wissenschaft. Nordisk Statistik Tidskrift, Bd. 1, 1922, S. 39 u. f.)

Erster Abschnitt.

Theorie der festen Merkmale.

§ 1. Bezeichnung der Merkmale und ihrer Verbindungen.

1. Wir wollen bei Kollektiven, deren Glieder nur darnach unterschieden werden, ob sie so oder so beschaffen, also nicht so groß, so schwer oder ähnliches sind, von alternativer Veränderlichkeit oder alternativer Variabilität sprechen.

Bei der Untersuchung solcher Kollektive sind zwei Prozesse zu unterscheiden: die Bildung der Klassen und die Auszählung der Klassen. Das Ergebnis kann als die Verteilung der Merkmale auf die Glieder des Kollektivs bezeichnet werden.

Ein Kollektiv umfaßt nur ganz ausnahmsweise alle Gegenstände der betreffenden Art, die zur Zeit seiner Bildung vorhanden waren. In der Regel bildet es nur einen ganz kleinen Teil davon, eine Stichprobe. Wenn man nun annimmt, daß die in dem Kollektiv angetroffene Verteilung der Merkmale ein getreues Abbild ihrer Verteilung in der Gesamtheit der betreffenden Gegenstände sei, so ist damit eine Hypothese aufgestellt, die sich im allgemeinen mit den Tatsachen nicht decken wird. Man wird auf Grund vielfältiger Erfahrung den Grundsatz aufstellen können, daß die Hypothese um so besser gestützt sei, je größer der Umfang des Kollektivs ist.

Doch spielt dabei die Art der Bildung des Kollektivs auch eine Rolle. Sie muß so vor sich gehen, daß keine Bevorzugung gewisser Merkmale oder Merkmalgruppen platzgreifen oder wenigstens keine auf eine solche Bevorzugung gerichtete Absicht verwirklicht werden kann. Man benützt, um dies zu erreichen, besondere Vorsichten, z. B. das wahllose oder willkürliche Herausgreifen der Individuen, die zum Kollektiv vereinigt werden sollen; die wahllose Numerierung und darauf das Herausheben nach einer festen Regel, etwa jedes 5., jedes 10. oder anderen Individuums, um jede Absichtlichkeit zu verhindern.

Es gibt aber auch Fälle, wo das Kollektiv in dem Sinne vollständig ist, daß es alle Individuen oder Exemplare umfaßt, die bei der Untersuchung in Betracht kommen können.

Will man z. B. die Verteilung der verschiedenen Haarfarben in der jugendlichen Bevölkerung eines Landes, einer Stadt feststellen, so wird man dazu eine Stichprobe[1]) verwenden, bestehend etwa in den schulbesuchenden Kindern eines

[1]) Über das Stichprobenverfahren hat R. Meerwarth in den beiden Aufsätzen „Über die repräsentative Methode“ (Zeitschrift des Preußischen Statistischen Landesamts, 72. Jahrgang, S. 352 u. f.) und „Beiträge zur repräsentativen Methode“ (Revue de l'Institut International de Statistique 1934, 4 p. 1 u. f.) grundlegende Forschungen angestellt. Meerwarth knüpft an eine Entschließung des Internationalen Statistischen Instituts an, in der gesagt wird, daß der als Stichprobe aus der Gesamtheit herausgenommene Bruchteil genügend repräsentativ für diese Gesamtheit sein muß, damit die Ergebnisse der Teilerhebung mit Berechtigung verallgemeinert werden können. Für die Entnahme der Stichprobe werden zwei

Bezirkes. Eine Volkszählung hingegen stellt die Verteilung bestimmter Merkmale, auf die man aus irgend welchen Gründen Wert legt, in der ganzen Bevölkerung, also in einem vollständigen Kollektiv fest.

Von den beiden erwähnten Prozessen erfordert nur die Klassenbildung eine besondere Erörterung. Sie ist ein wesentlich kombinatorisches Verfahren.

2. Wenn es sich nur um ein Merkmal handelt, das an den Gliedern des Kollektivs vorkommen oder auch fehlen kann, so sei das Merkmal und zugleich die Bejahung seines Vorkommens mit einem großen lateinischen Buchstaben, etwa A, sein Fehlen mit dem entsprechenden kleinen griechischen Buchstaben, also α, bezeichnet. Das Kollektiv zerfällt in zwei Klassen, kurz gesprochen die Klasse A und die Klasse α; erstere die positive, letztere die negative genannt; ihre Umfänge seien mit (A), (α) bezeichnet. Diesen einfachsten Fall der Klassenbildung nennt man Dichotomie.

Wird der Umfang des Kollektivs ein für allemal mit N bezeichnet, so hat man

$$N = (A) + (\alpha). \tag{1}$$

Sind zwei Merkmale vorhanden, die vorkommen und fehlen können, so mögen die Buchstaben A, B; α, β, in gleichem Sinne verwendet werden wie vorhin. Richtet man die Aufmerksamkeit auf beide Merkmale zugleich, so entstehen folgende Klassen:

$$AB, \quad \alpha B, \quad A\beta, \quad \alpha\beta, \tag{2}$$

wobei die Nebeneinanderstellung der Buchstaben das gleichzeitige Bestehen der durch sie angedeuteten Zustände ausdrückt: die Klasse AB umfaßt Individuen, an denen sowohl das Merkmal A als auch das Merkmal B auftritt; die Klasse αB vereinigt alle Individuen, welche das Merkmal B besitzen, an denen aber das Merkmal A fehlt usw.

Die angegebenen Klassen sind nach der Zahl der vereinigten Symbole Klassen 2. Ordnung, und das ist die höchste Ordnung, die bei zwei Merkmalen gebildet werden kann. Man kann in diesem Falle auch Klassen 1. Ordnung aufstellen, nämlich

$$A, \quad B, \quad \alpha, \quad \beta, \tag{3}$$

wenn man den Symbolen inklusiven Charakter zuerkennt, unter A also alle Individuen begreift, welchen die Eigenschaft A zukommt, gleichgültig, ob sie mit B oder β einhergeht u. s. w. Bei solcher Auffassung bestehen zwischen den

Hauptfälle unterschieden: die zufällige und die bewußte Auswahl. Meerwarth steht dem Prinzip der bewußten Auswahl skeptisch gegenüber und erwartet den Fortschritt von der Methode der zufälligen Auswahl.

W. Grävell kommt in dem Aufsatz „Die repräsentative Methode" (Deutsches Statistisches Zentralblatt 1923, S. 5 u. f.) zu dem Ergebnis, daß die repräsentative Methode in Zukunft vielleicht das beweglichste und billigste Handwerkzeug der amtlichen Statistik sein dürfte und werden müßte. Ihm scheint z. B. in der Kriminalstatistik, vielleicht auch in der Schulstatistik, der Landwirtschaftsstatistik oder in der Steuerstatistik ein weites Anwendungsfeld zu liegen.

Das Statistische Reichsamt führt in verschiedenen Gewerbezweigen Lohnerhebungen nach dem Prinzip der bewußten Auswahl durch. (Vgl. Wirtschaft und Statistik 1928, S. 163 u. f.)

Umfängen der Klassen, die wieder durch Einklammerung der Klassensymbole angedeutet werden sollen, die folgenden Beziehungen:

$$(A) = (AB) + (A\beta) \qquad (B) = (AB) + (\alpha B)$$
$$(\alpha) = (\alpha B) + (\alpha\beta) \qquad (\beta) = (A\beta) + (\alpha\beta). \tag{4}$$

$$N = (AB) + (\alpha B) + (A\beta) + (\alpha\beta). \tag{5}$$

Bei drei Merkmalen, die vorkommen und fehlen können, lassen sich Klassen aller Ordnungen von der dritten bis zur ersten bilden, je nachdem man die Aufmerksamkeit auf alle drei Merkmale zugleich oder nur auf zwei von ihnen oder nur auf eines richtet.

Die Klassen 3. Ordnung sind:

$$\begin{array}{ll} ABC & A\beta\gamma \\ \alpha BC & \alpha B\gamma \\ A\beta C & \alpha\beta C \\ AB\gamma & \alpha\beta\gamma; \end{array}$$

die Klassen 2. Ordnung:

$$\begin{array}{lll} AB & AC & BC \\ \alpha B & \alpha C & \beta C \\ A\beta & A\gamma & B\gamma \\ \alpha\beta & \alpha\gamma & \beta\gamma; \end{array}$$

die Klassen 1. Ordnung:

$$\begin{array}{lll} A & B & C \\ \alpha & \beta & \gamma. \end{array}$$

Bei dem inklusiven Charakter dieser Bezeichnungsweise läßt sich der Umfang jeder Klasse niederer Ordnung durch die Umfänge der Klassen der nächsthöheren Ordnung und daher schließlich durch die Umfänge der Klassen der höchsten Ordnung ausdrücken. Darum spielen die Klassen der höchsten Ordnung oder die letzten Klassen insofern eine bevorzugte Rolle, als durch ihre Umfänge die ganze Verteilung der Merkmale dargestellt werden kann. So hat man folgeweise

$$\begin{aligned} (A) &= (AB) + (A\beta) \\ (AB) &= (ABC) + (AB\gamma) \\ (A\beta) &= (A\beta C) + (A\beta\gamma), \end{aligned}$$

daher schließlich

$$(A) = (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma) \text{ u. s. w.}$$

Daraus folgt unmittelbar

$$(\alpha) = (\alpha BC) + (\alpha B\gamma) + (\alpha\beta C) + (\alpha\beta\gamma),$$

woraus sich weiter

$$(A) + (\alpha) =$$
$$= (ABC) + (\alpha BC) + (A\beta C) + (AB\gamma) + (A\beta\gamma) + (\alpha B\gamma) + (\alpha\beta C) + (\alpha\beta\gamma) = N$$

ergibt; ebenso ist

$$(B) + (\beta) = (C) + (\gamma) = N.$$

3. Versteht man unter positiven Klassen diejenigen, welche nur aus positiven Symbolen bestehen und zählt man zu ihnen als Klasse 0ter Ordnung das ganze

(ungegliederte) Kollektiv mit dem Umfang N, so besitzt man damit das Material, um die Umfänge aller übrigen Klassen zu berechnen.

Im Falle eines Merkmals hat man mit (A) und N das noch fehlende

$$(\alpha) = N - (A).$$

Im Falle zweier Merkmale bestimmt sich aus (AB), (A), (B) und N alles übrige; z. B. erhält man (αB) wie folgt:

$$(B) = (AB) + (\alpha B),$$

mithin

$$(\alpha B) = (B) - (AB);$$

und $(\alpha\beta)$

$$N = (AB) + (\alpha B) + (A\beta) + (\alpha\beta) = (AB) + (B) - (AB) + (A) - (AB) + (\alpha\beta),$$

mithin

$$(\alpha\beta) = N - (A) - (B) + (AB). \tag{6}$$

Im Falle dreier Merkmale genügen die Klassenumfänge (ABC), (AB), (AC), (BC), (A), (B), (C), N, um alles übrige zu finden; so erhält man z. B. aus

$$(B) = (AB) + (\alpha B)$$

durch Umstellung

$$(\alpha B) = (B) - (AB);$$

aus

$$(BC) = (ABC) + (\alpha BC)$$

ergibt sich

$$(\alpha BC) = (BC) - (ABC);$$

aus

$$(\alpha B) = (\alpha BC) + (\alpha B\gamma) = (BC) - (ABC) + (\alpha B\gamma)$$

findet man

$$(\alpha B\gamma) = (B) - (AB) - (BC) + (ABC) \text{ u. s. w.}; \tag{7}$$

aus

$$(\alpha\gamma) = (\alpha B\gamma) + (\alpha\beta\gamma),$$

wenn man darin $(\alpha B\gamma)$ durch den vorausgehenden und $(\alpha\gamma)$ durch den nach Vorschrift von (6) gebildeten Ausdruck ersetzt, erhält man:

$$(\alpha\beta\gamma) = N - (A) - (B) - (C) + (AB) + (AC) + (BC) - (ABC). \tag{8}$$

4. Es handelt sich noch darum, die Anzahl der Klassen einer bestimmten Ordnung und die Anzahl aller positiven Klassen bei n Merkmalen anzugeben, um einen allgemeinen Einblick in die Zusammenhänge zu erhalten.

Die Zahl der Klassen 1. Ordnung ist $2\binom{n}{1}$

" " " " 2. " " $2^2\binom{n}{2}$

Die Zahl der Klassen n. Ordnung ist $2^n\binom{n}{n}$,

weil jeder Platz in jeder Kombination der betreffenden Klasse auf zwei Arten, durch den lateinischen und durch den griechischen Buchstaben besetzt werden kann.

Die Zahl der **positiven** Klassen einer bestimmten Ordnung entsteht durch Weglassung des ersten Faktors, infolgedessen ist die Gesamtzahl der positiven Klassen, wenn man das ganze Kollektiv als den einzigen Vertreter der 0ten Klasse hinzunimmt,

$$1+\binom{n}{1}+\binom{n}{2}+\ldots+\binom{n}{n}=(1+1)^n=2^n.$$

So hat man also bei drei Merkmalen

1	Klasse	0ter	Ordnung
6	Klassen	1 „	„
12	„	2 „	„
8	„	3 „	„
8	positive	Klassen.	

Als wichtig ist hervorzuheben, daß die Anzahl der letzten Klassen übereinstimmt mit der Anzahl der positiven Klassen, beide Anzahlen betragen bei n Merkmalen 2^n, und beide Klassensysteme sind voneinander unabhängig, d. h. die Häufigkeit irgend einer letzten, bzw. einer positiven Klasse läßt sich durch die andern letzten, bzw. positiven Klassen nicht ausdrücken.

Der Umfang irgend einer Klasse $m$ter Ordnung hingegen läßt sich durch die Umfänge aller positiven Klassen dieser und der niedrigeren Ordnungen darstellen. Daraus folgt, daß die Umfänge der Klassen einer bestimmten Ordnung nicht unabhängig voneinander sind; vielmehr gibt es nur so viele unabhängige unter ihnen, als es positive Klassen der betreffenden und aller niedrigeren Ordnungen gibt, d. i. also, wenn m die Ordnung,

$$\binom{n}{m}+\binom{n}{m-1}+\binom{n}{m-2}+\cdots+1.$$

In entwickelter Form gibt dies

für $m=1$ $\quad n+1$

„ $m=2$ $\quad \frac{1}{2}(n^2+n+2)$

„ $m=3$ $\quad \frac{1}{6}(n^3+5n+6)$

„ $m=4$ $\quad \frac{1}{24}(n^4-2n^3+11n^2+14n+24)$

„ $m=5$ $\quad \frac{1}{120}(n^5-5n^4+25n^3+5n^2+94n+120)$ usw.

Hiernach gibt es beispielsweise bei 4 Merkmalen (Elementen)

$2\binom{4}{1}=$ 8 Klassen 1. Ordnung, davon aber nur 5 voneinander unabhängig;

$2^2\binom{4}{2}=$ 24 Klassen 2. Ordnung, davon nur 11 voneinander unabhängig;

$2^3\binom{4}{3}=$ 32 Klassen 3. Ordnung, davon 15 unabhängig; schließlich

$2^4\binom{4}{4}=$ 16 Klassen 4. Ordnung, die untereinander unabhängig sind.

5. Beispiel. Die Geburten werden nach drei Merkmalen klassifiziert:

$$A \text{ lebend}, \quad B \text{ ehelich}, \quad C \text{ männlich},$$

die entsprechenden negativen Merkmale sind:

$$\alpha \text{ tot}, \quad \beta \text{ unehelich}, \quad \gamma \text{ weiblich}.$$

Die im Deutschen Reich 1933[1]) verzeichneten Geburten ergaben in den letzten Klassen folgende Häufigkeitszahlen:

$$\begin{array}{llll} (ABC) = 441\,135 & (A\beta C) = 52\,338 & (\alpha BC) = 13\,462 & (\alpha\beta C) = 2\,292 \\ (AB\gamma) = 413\,719 & (A\beta\gamma) = 49\,779 & (\alpha B\gamma) = 10\,513 & (\alpha\beta\gamma) = 1\,829; \end{array}$$

als Gesamtsumme folgt daraus

$$N = 985\,067.$$

Die positiven Klassen erhält man nach den Schemata

$$(AB) = (ABC) + (AB\gamma), \quad (AC) = (ABC) + (A\beta C), \quad (BC) = (ABC) + (\alpha BC),$$

$$\begin{aligned} (A) &= (ABC) + (A\beta C) + (AB\gamma) + (A\beta\gamma) \\ (B) &= (ABC) + (\alpha BC) + (AB\gamma) + (\alpha B\gamma) \\ (C) &= (ABC) + (\alpha BC) + (A\beta C) + (\alpha\beta C); \end{aligned}$$

mithin ist die ganze Verteilung auch durch das folgende Zahlensystem dargestellt oder bestimmt:

N	$= 985\,067$	alle Geborenen
(A)	$= 956\,971$	lebend geborene Kinder
(B)	$= 878\,829$	ehelich geborene Kinder
(C)	$= 509\,227$	Knaben
(AB)	$= 854\,854$	lebend geborene eheliche Kinder
(AC)	$= 493\,473$	lebend geborene Knaben
(BC)	$= 454\,597$	ehelich geborene Knaben
(ABC)	$= 441\,135$	lebend geborene eheliche Knaben.

In der Tat kann man aus diesem System die Häufigkeit jeder letzten Klasse wiedergewinnen, z. B.

$(\alpha B\gamma) = (B) - (AB) - (BC) + (ABC) = 10513$ totgeborene eheliche Mädchen,

$(\alpha\beta\gamma) = N - (A) - (B) - (C) + (AB) + (AC) + (BC) - (ABC) = 1829$ totgeborene uneheliche Mädchen.

6. Es gibt, wie gezeigt worden ist, zwei ausgezeichnete Systeme von Klassenhäufigkeiten, die zur vollständigen Bestimmung der Merkmalverteilung in einem Kollektiv ausreichen: das System der letzten Klassen und das System der positiven Klassen, beide 2^n Zahlen umfassend, wenn n Merkmale im Spiele sind. Auch jedes andere gemischte System gleichen Umfangs würde dazu geeignet sein: doch sind die zwei genannten die praktisch üblichen.

Auf die Frage, ob jedes System von 2^n Zahlwerten eine wirklich mögliche Verteilung darstellen kann, fällt die Antwort verschieden aus, je nachdem die gegebenen Werte die letzten Klassen oder die positiven Klassen sind.

[1]) Statistisches Jahrbuch für das Deutsche Reich 1935, S. 39.

Jede Klassenhäufigkeit kann nur eine positive Zahl oder Null sein; negative Zahlen sind ausgeschlossen. Nun setzen sich die Häufigkeitszahlen der positiven Klassen aus den Häufigkeitszahlen der letzten Klassen nur durch Addition zusammen (siehe die ersten zwei Gleichungen (4) und die Gleichung (5)).

Ist demnach ein System von (positiven) Häufigkeitszahlen der letzten Klassen gegeben, so ergeben sich auch für die positiven Klassen positive Zahlen: jedes so beschaffene System kann also eine wirkliche Verteilung darstellen.

Anders steht es mit den Häufigkeitszahlen der letzten Klassen; diese setzen sich aus den Häufigkeitszahlen der positiven Klassen durch Addition und Subtraktion zusammen, und nur wenn solche Bedingungen zwischen den letzteren Häufigkeitszahlen erfüllt sind, daß auch alle Häufigkeitszahlen der Endklassen positiv (nicht negativ) ausfallen, kann das gegebene System einer wirklichen Verteilung entsprechen.

Sind also die Häufigkeitszahlen der positiven Klassen, selbstverständlich als positive Zahlen, gegeben, so hat man, um über die Frage ihrer Verträglichkeit zu entscheiden, die Häufigkeitszahlen aller Endklassen abzuleiten, und nur wenn diese sämtlich positiv (oder gleich Null) ausfallen, besteht Verträglichkeit.

Eine solche Untersuchung ist in jedem derartigen Falle zu Kontrollzwecken notwendig, damit nicht Zahlen als Darstellung einer wirklichen Verteilung hingenommen werden, die es gar nicht sein können.

Die Anwendung der im vorstehenden entwickelten Gleichungen erleichtert die planmäßige Auffindung sämtlicher Kontrollrechnungen, die sich bei den Tabellen in der praktischen Statistik notwendig machen.

§ 2. Abhängigkeit von Merkmalen.

7. Zwei Merkmale A, B, die auf die Gegenstände eines Kollektivs verteilt sind, können aufeinander eine Anziehung ausüben in dem Sinne, daß es in der Natur dieser Gegenstände liegt, das Merkmal A leichter mit dem Merkmal B auftreten zu lassen als mit seiner Negation (seinem Gegensatze) β; es kann aber zwischen beiden Merkmalen auch eine Abstoßung obwalten in dem Sinne, daß das Merkmal A eher mit β als mit B verbunden erscheint. Damit aber diese Unterscheidung verständlich sei, ist es notwendig, die Indifferenz der beiden Merkmale zu definieren.

Als indifferent sollen die beiden Merkmale dann erklärt werden, wenn sich das Merkmal A auf die Gegenstände des Merkmals B mit derselben relativen Häufigkeit verteilt wie auf die Gegenstände des Merkmals β.

Demnach ist das Kennzeichen der Indifferenz ausgedrückt durch den Ansatz

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} \tag{1}$$

Hiermit ist es nun möglich, auch die beiden erstgedachten Fälle arithmetisch zu kennzeichnen:

$$\frac{(AB)}{(B)} > \frac{(A\beta)}{(\beta)} \tag{2}$$

bedeutet die Anziehung zwischen A und B,

$$\frac{(AB)}{(B)} < \frac{(A\beta)}{(\beta)} \tag{3}$$

die Abstoßung. Eine andere Ausdrucksweise ist es, wenn man die Merkmale im Falle (1) als voneinander **unabhängig**, in den Fällen (2), (3) als voneinander **abhängig** bezeichnet und die Fälle (2) und (3) als **positive** und **negative** Abhängigkeit unterscheidet, entsprechend den Vorzeichen der Differenz $\frac{(AB)}{(B)} - \frac{(A\beta)}{(\beta)}$[1]).

Wenn die Bedingung der Unabhängigkeit, bzw. der Abhängigkeit für ein Merkmalpaar besteht, so gilt sie notwendig auch für die anderen Paare des Merkmalkomplexes, d. h. aus (1) folgt notwendig

$$\frac{(BA)}{(A)} = \frac{(B\alpha)}{(\alpha)}$$

$$\frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)}$$

$$\frac{(\beta A)}{(A)} = \frac{(\beta\alpha)}{(\alpha)}.$$

Es ergibt sich nämlich aus (1)

$$\frac{(B)-(AB)}{(B)} = \frac{(\beta)-(A\beta)}{(\beta)}$$

d. h.

$$\frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)}.$$

womit die zweite Gleichung der Gruppe erwiesen ist. Aus (1) und der letztgewonnenen Gleichung folgt weiter

$$\frac{(AB)+(A\beta)}{(B)+(\beta)} = \frac{(A)}{N} = \frac{(AB)}{(B)}$$

$$\frac{(\alpha B)+(\alpha\beta)}{(B)+(\beta)} = \frac{(\alpha)}{N} = \frac{(\alpha B)}{(B)}$$

und daraus weiter

$$\frac{(BA)}{(A)} = \frac{(B\alpha)}{(\alpha)},$$

also die erste der zu beweisenden Gleichungen; die dritte erhält man aus dieser auf dem erstbeschrittenen Wege.

Zugleich hat sich im Laufe der Ableitung die Gleichung

$$\frac{(AB)}{(B)} = \frac{(A)}{N}$$

ergeben, wofür auch

$$\frac{(AB)}{(A)} = \frac{(B)}{N}$$

[1]) Gini gebraucht für die drei Verhaltungsweisen der Merkmale A, B, die hier als positive Abhängigkeit, Unabhängigkeit und negative Abhängigkeit benannt worden sind, die Bezeichnungen: Konkordanz, Indifferenz, Diskordanz. Sul criterio di concordanza tra due caratteri — und Indici di concordanza (Atti del Reale Ist. Veneto, 1915—1916, t. LXXV).

geschrieben werden kann; aus beiden folgt

$$(AB) = \frac{(A)(B)}{N}$$

und schließlich

$$\frac{(AB)}{N} = \frac{(A)}{N}\,\frac{(B)}{N}. \tag{4}$$

Gerade diese letzte Form ist wegen ihrer Beziehung zur Wahrscheinlichkeitsrechnung bemerkenswert. Sie besagt, daß die relative Häufigkeit der Merkmalverbindung AB gleich ist dem Produkte der relativen Häufigkeiten der einzelnen Merkmale, wofern die Merkmale voneinander unabhängig sind. Dieser Satz gilt, wie sich aus den obigen Ansätzen erweisen läßt, für alle Merkmalpaare aus dem Komplex, also ist auch

$$\frac{(A\beta)}{N} = \frac{(A)}{N}\,\frac{(\beta)}{N},$$

$$\frac{(\alpha\beta)}{N} = \frac{(\alpha)}{N}\,\frac{(\beta)}{N}$$

usw.; daraus geht auch hervor, daß für unabhängige Merkmale die Beziehung stattfindet:

$$(AB)(\alpha\beta) = (\alpha B)(A\beta) = \frac{(A)(B)(\alpha)(\beta)}{N^2}. \tag{5}$$

8. Dies alles waren Folgerungen der Gleichung (1) und der Beziehungen zwischen den Klassenhäufigkeiten. Tritt also an die Stelle von (1) eine der Ungleichungen (2), (3), so kann in keiner der aus (1) abgeleiteten Relationen mehr das Gleichheitszeichen herrschen, d. h. die Abhängigkeit ist ebenso wie die Unabhängigkeit etwas Gegenseitiges.

Die Bedingung (2) für positive Abhängigkeit zwischen A und B hat eine Reihe von Folgerungen zwischen den Häufigkeitszahlen, von denen jede kennzeichnend ist für die eben vorhandene Sachlage. Bringt man (2) auf die Form

$$(AB)(\beta) > (A\beta)(B)$$

und addiert hierzu die Identität

$$(AB)(B) = (AB)(B),$$

so ergibt sich

$$\frac{(AB)}{(B)} > \frac{(A)}{N} \quad \text{und} \quad \frac{(BA)}{(A)} > \frac{(B)}{N}. \tag{6}$$

Subtrahiert man

$$N(AB) > (A)(B)$$

von der Identität $N(A) = N(A)$, so kommt man zu

$$\frac{(\beta A)}{(A)} < \frac{(\beta)}{N} \quad \text{und} \quad \frac{(A\beta)}{(\beta)} < \frac{(A)}{N}. \tag{7}$$

In gleicher Weise entsteht durch Subtraktion derselben Ungleichung von $N(B) = N(B)$

$$\frac{(\alpha B)}{(B)} < \frac{(\alpha)}{N} \quad \text{und} \quad \frac{(B\alpha)}{(\alpha)} < \frac{(B)}{N}. \tag{8}$$

Nimmt man (7) zum Ausgangspunkt, so erhält man mit Hilfe der Identität $N(\beta) = N(\beta)$ durch analoge Überlegungen

$$\frac{(\alpha\beta)}{(\beta)} > \frac{(\alpha)}{N} \quad \text{und} \quad \frac{(\beta\alpha)}{(\alpha)} > \frac{(\beta)}{N} \tag{9}$$

Schreibt man dies in der Form

$$N(\alpha\beta) > (\alpha)(\beta),$$

ersetzt erstens (α) durch $(\alpha B) + (\alpha\beta)$, subtrahiert $(\alpha\beta)$ (β) und ersetzt zweitens (β) durch $(A\beta) + (\alpha\beta)$, so ergeben sich die Beziehungen

$$\frac{(\alpha\beta)}{(\beta)} > \frac{(\alpha B)}{(B)} \quad \text{und} \quad \frac{(\beta\alpha)}{(\alpha)} > \frac{(A\beta)}{(A)} \tag{10}$$

Kennzeichnen die Ungleichungen (2) und (6) — (10) in ihrer Gesamtheit den Fall einer positiven Abhängigkeit zwischen A und B, so braucht man in ihnen nur das Ungleichheitszeichen umzukehren, um die im Falle negativer Abhängigkeit zwischen A und B geltenden Ungleichungen zu erhalten.

Wiewohl nun jede der zahlreichen Beziehungen geeignet wäre, über eine vorhandene Abhängigkeit und ihre Richtung auszusagen, so wird doch eine Wahl zu treffen sein, die auf verschiedene Umstände Rücksicht zu nehmen hat. Allem voran muß die zu lösende Frage den Ausschlag geben: jene Beziehung ist allen andern vorzuziehen, die auf die gestellte Frage eine direkte Antwort gibt. Des weiteren wird es von dem zur Verfügung stehenden Beobachtungsmaterial abhängen, zu welchen Beziehungen man greift. Darüber lassen sich allgemeine Regeln nicht aufstellen.

Es liegt in der Natur der Sache, daß die Bedingungen der Unabhängigkeit, die sich in der Form von Gleichungen ergeben haben, kaum jemals in voller Strenge erfüllt sein werden, weil Unabhängigkeit, auch wenn sie wirklich vorhanden ist, nie in voller Reinheit zutage treten wird; vielmehr wird sie durch zufällige Störungen verwischt sein. Ob nun eine der Ungleichungen, die für Abhängigkeit kennzeichnend sind, nur durch zufällige Störungen entstanden ist oder die Folge einer wirklichen Abhängigkeit sei, darüber läßt sich nicht so ohneweiters ein Urteil abgeben. Ein Weg, zu einem solchen zu gelangen, besteht darin, daß man an einer Mehrzahl von Kollektiven, die entweder verschiedenen Orten oder verschiedenen Zeiten entstammen oder sonstwie in ihrer Abkunft voneinander abweichen, dieselbe Frage untersucht; zeigt sich immer dasselbe Verhalten, dann kann mit großer Berechtigung auf Abhängigkeit in einem bestimmten Sinne geschlossen werden; schwankt das Verhalten so, daß die Ungleichheit bald in der einen, bald in der anderen Richtung stattfindet, dann ist zu vermuten, daß man es mit bloßen Zufallsschwankungen und nicht mit Abhängigkeit zu tun hat. Man wird also zwischen deutlich ausgesprochenen und zweifelhaften Fällen zu unterscheiden haben. Diese Unterscheidung hängt bereits mit einer Vorstellung zusammen, auf die alsbald eingegangen werden soll, mit der Vorstellung nämlich, daß die Abhängigkeit etwas Steigerungsfähiges, Graduelles sei. Vorher möge an einigen Beispielen die bloße Frage nach dem Vorhandensein einer Abhängigkeit geprüft werden.

9. Erstes Beispiel. Besteht zwischen Geschlecht und Lebendgeburt eine Abhängigkeit?

Bezeichnet A die Lebend-, also α die Totgeburt, B das männliche, β das weibliche Geschlecht, so ergibt die deutsche Statistik aus 1933[1]) folgende Klassenhäufigkeiten:

$$(AB) = 1065 \qquad (A\beta) = 1000$$
$$(\alpha B) = 1276 \qquad (\alpha\beta) = 1000;$$

diese Statistik führt nämlich an, wieviel Knaben auf je 1000 Mädchen der betreffenden Kategorie entfallen; das Kollektiv, auf das sich diese Daten beziehen, ist also gebildet: aus 1000 lebendgeborenen Mädchen und den auf sie entfallenden Knaben und 1000 totgeborenen Mädchen und den auf sie entfallenden Knaben, zusammen aus $N = 4341$ Individuen. Die Besetzung der Klassen erster Ordnung ist die folgende:

$$(A) = 2065 \qquad (B) = 2341 \qquad (\alpha) = 2276 \qquad (\beta) = 2000.$$

Daraus berechnet sich:

$$\frac{(BA)}{(A)} = \frac{1065}{2065} = 0{,}516, \qquad \frac{(B\alpha)}{(\alpha)} = \frac{1276}{2276} = 0{,}561;$$

es ist also $\frac{(BA)}{(A)} < \frac{(B\alpha)}{(\alpha)}$, somit besteht im Sinne von (3) zwischen männlichem Geschlecht und Lebendgeburt eine negative Abhängigkeit. In Worten ausgedrückt heißt dies: unter den Lebendgeborenen sind Knaben weniger häufig als unter den Totgeborenen.

In einer andern Weise:

$$\frac{(A\beta)}{(\beta)} = \frac{1000}{2000} = 0{,}5, \qquad \frac{(A)}{N} \quad \frac{2065}{4341} = 0{,}476,$$

also ist $\frac{(A\beta)}{(\beta)} > \frac{(A)}{N}$, was nach (7) wieder im Einklang steht mit einer negativen Abhängigkeit des männlichen Geschlechts von der Lebendgeburt und besagt, daß unter den weiblichen Geburten die Lebendgeburten häufiger sind als in der Gesamtheit.

In noch anderer Art:

$$\frac{(\alpha\beta)}{(\beta)} = \frac{1000}{2000} = 0{,}5, \qquad \frac{(\alpha B)}{(B)} = \frac{1276}{2341} = 0{,}545,$$

mithin $\frac{(\alpha\beta)}{(\beta)} < \frac{(\alpha B)}{(B)}$, was zufolge (10) zu dem gleichen Schlusse führt und ausdrückt, daß unter den weiblichen Geburten die Totgeburten minder häufig sind als unter den männlichen.

Die erste und dritte Art lassen die Abhängigkeit ziffermäßig schärfer hervortreten als die zweite Art.

10. Zweites Beispiel. Lebendgeburt und Ehelichkeit.

Bezeichnet man diese beiden Tatsachen mit A und B, so bedeuten α, β Totgeburt und Unehelichkeit. Die deutsche Statistik für 1933[2]) weist folgende Klassenhäufigkeiten auf:

[1]) Statistisches Jahrbuch für das Deutsche Reich 1935, S. 39.
[2]) Statistisches Jahrbuch für das Deutsche Reich 1935, S. 39.

$$(AB) = 131, \qquad (\alpha B) = 4, \qquad (A\beta) = 16, \qquad (\alpha\beta) = 1;$$

daraus berechnen sich

$$N = 152; \quad (A) = 147, \quad (B) = 135, \quad (\alpha) = 5, \quad (\beta) = 17.$$

Die Zahlen beziehen sich auf 10 000 Einwohner.

Man findet

$$\frac{(AB)}{(B)} = \frac{131}{135} = 0{,}970, \qquad \frac{(A\beta)}{(\beta)} = \frac{16}{17} = 0{,}941;$$

es besteht also zwischen dem ersten und dem zweiten Merkmal eine sehr schwache positive Abhängigkeit, soweit man nach diesen Zahlen urteilen kann. Etwas stärker tritt sie hervor, wenn man

$$\frac{(B\alpha)}{(\alpha)} = \frac{4}{5} = 0{,}8 \text{ mit } \frac{(B)}{N} = \frac{135}{152} = 0{,}888$$

vergleicht; es ist, wie es zufolge (8) sein soll, $\frac{(B\alpha)}{(\alpha)} < \frac{(B)}{N}$.

11. Drittes Beispiel. Taubstummheit und geistige Gebrechlichkeit. Bezeichnet A geistige Gebrechlichkeit, B Taubstummheit, so ergab eine Statistik aus den Jahren 1925/26 für das Deutsche Reich[1]) folgende Zahlen:

N	Gesamtbevölkerung	62 410 619
(A)	Zahl der Geistig-Gebrechlichen	230 112
(B)	Zahl der Taubstummen	45 376
(AB)	Zahl der taubstummen Geistig-Gebrechlichen	2 613

Bei dieser Sachlage eignen sich die Beziehungen (6) zur Entscheidung der Frage, ob zwischen geistiger Gebrechlichkeit und Taubstummheit eine Abhängigkeit besteht.

$$\frac{(AB)}{(B)} = \frac{2613}{45376} = 0{,}0576, \qquad \frac{(A)}{N} = \frac{230112}{62410619} = 0{,}0037,$$

$$\frac{(BA)}{(A)} = \frac{2613}{230112} = 0{,}0114, \qquad \frac{(B)}{N} = \frac{45376}{62410619} = 0{,}0007.$$

Übereinstimmend ist also

$$\frac{(AB)}{(B)} > \frac{(A)}{N} \quad \text{und} \quad \frac{(BA)}{(A)} > \frac{(B)}{N}$$

und der Unterschied ist beidemal sehr erheblich, so daß von einer ausgesprochenen positiven Abhängigkeit zwischen Taubstummheit und geistiger Gebrechlichkeit gesprochen werden kann. Die Ansätze besagen folgendes: Während die Geistig-Gebrechlichen unter den Taubstummen 57,6‰ ausmachen, sind sie in der Gesamtbevölkerung nur mit 3,7‰ vertreten, und während die Taubstummen unter den Geistig-Gebrechlichen mit 11,4‰ vorkommen, betragen sie von der ganzen Bevölkerung nur 0,7‰.

---

[1]) Statistisches Jahrbuch für das Deutsche Reich 1932, S. 408.

12. Viertes Beispiel. Augenfarbe von Vater und Sohn.

Unterscheidet man hellfarbige und dunkle Augen, wobei zu den ersteren die blauen und grauen, zu den letzteren die braunen und schwarzen Augen zählen, und bezeichnet lichte und dunkle Farbe beim Vater mit A, α, beim Sohn mit B, β, so ergab eine von F. Galton[1]) gepflogene Erhebung, bei der jeder Vater so oft gezählt wurde, als er Söhne hatte, folgende Statistik:

$$(AB) = 471, \qquad (A\beta) = 151, \qquad (\alpha B) = 148, \qquad (\alpha\beta) = 230;$$

man berechnet daraus

$$N = 1000; \quad (A) = 622, \quad (B) = 619, \quad (\alpha) = 378, \quad (\beta) = 381.$$

In Prozenten ausgedrückt ergeben sich die Quotienten

$$\frac{(BA)}{(A)} = \frac{47100}{622} = 76\%, \qquad \frac{(B\alpha)}{(\alpha)} = \frac{14800}{378} = 39\%, \text{ also } \frac{(BA)}{(A)} > \frac{(\alpha B)}{(\alpha)};$$

d. h. 76% der helläugigen Väter haben lichtäugige Söhne, aber nur 39% der dunkeläugigen Väter lichtäugige Söhne.

Es besteht also eine deutliche positive Abhängigkeit zwischen der Augenfarbe der Söhne und jener der Väter. Als nicht sachgemäß ist es zu bezeichnen, umgekehrt von der Augenfarbe des Sohnes auf die des Vaters zu schließen.

13. Als Norm für die Beurteilung der Abhängigkeit zweier Merkmale A, B soll jenes Verhalten dienen, das bei ihrer Unabhängigkeit Geltung hätte; je mehr sich das wirkliche Verhalten von diesem letzteren entfernt, für um so stärker wird man die Abhängigkeit erklären.

Nun stellen sich die Klassenhäufigkeiten zweiter Ordnung (AB), $(A\beta)$, (αB), $(\alpha\beta)$, wenn Unabhängigkeit besteht, zufolge (4) durch die Klassenhäufigkeiten erster Ordnung wie folgt dar:

$$\frac{(A)(B)}{N}, \qquad \frac{(A)(\beta)}{N}, \qquad \frac{(\alpha)(B)}{N}, \qquad \frac{(\alpha)(\beta)}{N}.$$

Bezeichnet also (AB) die wirklich beobachtete, $[AB]$ die nach der vorstehenden Formel berechnete Klassenhäufigkeit, so werden diese Größen nur bei strenger Unabhängigkeit übereinstimmen, bei Abhängigkeit aber voneinander abweichen, und die Größe dieser Abweichung

$$(AB) - [AB] = \delta \tag{11}$$

ist eine die Stärke der Abhängigkeit kennzeichnende Größe.

Da sowohl $(AB) + (A\beta) = (A)$ als auch $[AB] + [A\beta] = (A)$ ist, so folgt daraus

$$(AB) - [AB] + (A\beta) - [A\beta] = 0,$$

somit ist

$$(A\beta) - [A\beta] = -\delta. \tag{12}$$

[1]) Vgl. K. Pearson, Philosophical Transactions of the Royal Society of London. A, vol. 195 (1901), p. 138.

In gleicher Weise ergibt sich aus der Tatsache, daß $(AB) + (\alpha B) = (B)$ und ebenso $[AB] + [\alpha B] = (B)$, daß auch

$$(\alpha B) - [\alpha B] = -\delta. \tag{13}$$

Endlich führen die Gleichungen $(\alpha B) + (\alpha\beta) = (\alpha)$ und $[\alpha B] + [\alpha\beta] = (\alpha)$ dazu, daß

$$(\alpha\beta) - [\alpha\beta] = \delta \tag{14}$$

ist.

Die Abhängigkeit zwischen A und B ist also gleichgerichtet mit der zwischen α und β, ebenso sind die Abhängigkeiten zwischen A und β einerseits und α und B anderseits untereinander gleichgerichtet, den erstgenannten gegenüber aber entgegengesetzt. Bei $\delta > 0$ ist die Abhängigkeit der betreffenden Merkmale positiv, bei $\delta < 0$ negativ.

Ersetzt man in (11) $[AB]$ durch seinen Ausdruck und führt durchwegs Klassenhäufigkeiten zweiter Ordnung ein, so wird

$$\left.\begin{aligned}\delta &= (AB) - \frac{(A)(B)}{N}\\ &= \frac{1}{N}\left\{[(AB)+(A\beta)+(\alpha B)+(\alpha\beta)](AB)-[(AB)+(A\beta)][(AB)+(\alpha B)]\right\}\\ &= \frac{1}{N}\left\{(AB)(\alpha\beta) - (A\beta)(\alpha B)\right\}\\ &= \frac{1}{N}\begin{vmatrix}(AB) & (A\beta)\\ (\alpha B) & (\alpha\beta)\end{vmatrix};\end{aligned}\right\} \tag{15}$$

demnach entscheidet das Vorzeichen der Determinante über die Richtung der Abhängigkeit und ihr absoluter Wert in Verbindung mit N über deren Stärke.

Will man $(A\beta) - [A\beta]$ haben, so braucht man in (15) bloß B mit β zu vertauschen; dadurch verwandelt sich die Determinante in

$$\begin{vmatrix}(A\beta) & (AB)\\ (\alpha\beta) & (\alpha B)\end{vmatrix},$$

d. h. sie ändert bloß ihr Vorzeichen, in Übereinstimmung mit dem vorhin Ausgeführten. Ebenso erfährt die Determinante bloß eine Umstellung ihrer Reihen, wenn man auf $(\alpha B) - [\alpha B]$ und $(\alpha\beta) - [\alpha\beta]$ ausgeht.

14. Die Größe δ ist ein absolutes Maß der Abhängigkeit und daher zu Vergleichen im allgemeinen minder geeignet.

Liegt das Bedürfnis nach einem relativen Maß vor, so muß man an dieses vor seiner Aufstellung gewisse in der Natur der Sache begründete Forderungen stellen.

Die erste geht dahin, daß dieses Maß im Falle der Unabhängigkeit einen bestimmten Wert annehme, und es erscheint am natürlichsten, wenn man diesen mit Null festsetzt, so daß die beiden Arten der Abhängigkeit sich im Vorzeichen aussprechen.

Ferner erscheint es zweckmäßig, den äußersten Grenzen der Abhängigkeit bestimmte Zahlwerte zuzuordnen. Die obere Grenze der Abhängigkeit zwischen

A und B ist dann erreicht, wenn entweder jedes A auch B oder jedes B auch A ist oder wenn beides zugleich stattfindet; es ist dann $(A\beta)$, bzw. (αB) gleich Null oder es findet beides zugleich statt. Unter diesen Umständen soll von vollständiger positiver Abhängigkeit gesprochen werden, und ihr Maß soll $+1$ sein. Als untere Grenze wird der Fall anzusehen sein, wenn A niemals mit B oder α niemals mit β zusammentrifft oder wenn beides zugleich stattfindet; dann ist entweder (AB) oder $(\alpha\beta)$ oder sind beide zugleich Null. Eine solche Sachlage möge als vollständige negative Abhängigkeit bezeichnet und ihr die Zahl -1 zugeordnet werden. Die verschiedenen Grade positiver und negativer Abhängigkeit sind dann durch positive und negative echte Brüche ausgedrückt.

Ein Ausdruck, der allen diesen Forderungen genügt, ist der von Yule[1]) eingeführte Abhängigkeitskoeffizient

$$q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)};$$

denn bei Unabhängigkeit ist der Zähler Null; bei vollständiger positiver Abhängigkeit wird $q = +1$; bei vollständiger negativer Abhängigkeit wird $q = -1$. Mit Berücksichtigung von (15) kann für q auch geschrieben werden

$$q = \frac{N\delta}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}, \tag{16}$$

wodurch die Zahl q mit dem absoluten Maß δ der Abhängigkeit in Zusammenhang gebracht ist.

15. Beispiele:

1) Augenfarbe bei Ehegatten.

Unter A die Lichtfarbigkeit der Augen des Gatten, unter B die Lichtfarbigkeit der Augen der Gattin verstanden, so daß α, β die Dunkelfarbigkeit bedeuten, hat eine von Francis Galton[2]) ausgeführte Statistik folgende Verteilung ergeben:

$$(AB) = 309, \qquad (A\beta) = 214, \qquad (\alpha B) = 132, \qquad (\alpha\beta) = 119.$$

Daraus leiten sich ab die Zahlen:

$$N = 774; \quad (A) = 523, \quad (B) = 441, \quad (\alpha) = 251, \quad (\beta) = 333,$$

aus welchen sich weiter berechnet:

$$[AB] = 298, \qquad [A\beta] = 225, \qquad [\alpha B] = 143, \qquad [\alpha\beta] = 108.$$

Die Zusammenstellung der beobachteten und der unter Voraussetzung der Unabhängigkeit gerechneten Häufigkeiten:

[1]) Yule, Philosophical Transactions of the Royal Society of London, A, vol. 194 (1900), p. 272.

[2]) Vgl. K. Pearson, Philosophical Transactions of the Royal Society of London, A, vol. 195 (1901), p. 141.

		Differenz
309	297,9884	11,0116
214	225,0116	−11,0116
132	143,0116	−11,0116
119	107,9884	11,0116

illustriert die theoretischen Entwicklungen und weist auf eine positive Abhängigkeit hin. Der Abhängigkeitskoeffizient berechnet sich mit

$$q = \frac{8523}{65019} = 0{,}131$$

und spricht gleichfalls dafür, daß eine, wenn auch schwache, Neigung zur Gleichfarbigkeit der Augen der Eheschließenden obwaltet. Die „Differenzen" schwanken dem Betrage nach zwischen 3,6 und 9,3 % der beobachteten Zahl.

2) Ehelichkeit und Geschlecht bei den Totgeburten.

Mit A sei die eheliche, mit B die männliche Totgeburt, mit α, β also die uneheliche, bzw. die weibliche bezeichnet.

Die deutsche Statistik für 1933[1]) gibt, auf 1000 Totgeburten gerechnet, folgende Besetzung der Klassen:

$$(AB) = 479{,}143, \quad (A\beta) = 374{,}181, \quad (\alpha B) = 81{,}578, \quad (\alpha\beta) = 65{,}098.$$

Man findet daraus

$$(A) = 853{,}324, \quad (B) = 560{,}721, \quad (\alpha) = 146{,}676, \quad (\beta) = 439{,}279;$$

$$[AB] = 478{,}477, \quad [A\beta] = 374{,}847, \quad [\alpha B] = 82{,}244, \quad [\alpha\beta] = 64{,}432;$$

die gemeinsame Differenz, vom Vorzeichen abgesehen, beträgt nur 0,666. Dementsprechend stellt sich auch der Abhängigkeitskoeffizient sehr klein heraus, nämlich

$$q = \frac{666}{61716} = 0{,}011,$$

so daß man von einer Abhängigkeit zwischen Ehelichkeit und Geschlecht bei den Totgeburten nicht mit Sicherheit sprechen kann; die Abweichungen mögen eher als zufällig anzusehen sein.

3) Darwin[2]) hat sich mit dem Einfluß der Abstammung, ob aus Kreuzung oder Selbstbefruchtung hervorgegangen, auf das Wachstum von Pflanzen beschäftigt und hat über eine große Anzahl von Arten und Varietäten Versuche nach dieser Richtung angestellt. Yule griff aus diesem Beobachtungsmaterial 38 Spezies heraus, die ihm eine genügende Grundlage für eine rechnerische Behandlung zu bieten schienen, im ganzen 1094 Pflanzen.

[1]) Statistisches Jahrbuch für das Deutsche Reich 1935, S. 39.

[2]) Vgl. G. U. Yule, Philosophical Transactions of the Royal Society of London, A, vol. 194 (1900), p. 293.

Gibt man der Kreuzung das Zeichen A, der Selbstbefruchtung das Zeichen α, der Hochwüchsigkeit das Zeichen B, dem niedrigen Wuchs das Zeichen β, so ist die erste dieser Dichotomien eine absolute, die andere aber eine relative, weil es einer Festsetzung bedarf, wann von Hochwuchs und wann von dem Gegenteil gesprochen werden soll. Als Beziehungsgröße wurde das arithmetische Mittel der Höhen der betreffenden Versuchspflanzen genommen, so daß nunmehr genauer B über Mittelgröße, β unter Mittelgröße bedeutet.

Das ganze Material zeigte folgende Gliederung:

$$(AB) = 395, \qquad (\alpha B) = 179, \qquad (A\beta) = 168, \qquad (\alpha\beta) = 372;$$

daraus ergeben sich die Zahlwerte

$$\mathfrak{d} = 104{,}908, \qquad q = 0{,}660;$$

es herrscht somit im großen ganzen zwischen Kreuzung und Hochwuchs eine deutliche positive Abhängigkeit.

Bei den einzelnen Arten erwies sie sich verschieden hoch, schlug bei einigen in negative Abhängigkeit um, und in einigen wenigen Fällen wiesen die Zahlen auf Unabhängigkeit. Wir greifen drei Spezies heraus, die positive Abhängigkeit in steigendem Grade zeigen, nämlich:

1. Lobelia fulgens: $(AB) = 17$, $(\alpha B) = 12$, $(A\beta) = 17$, $(\alpha\beta) = 22$;
2. Reseda odorata: $(AB) = 39$, $(\alpha B) = 25$, $(A\beta) = 16$, $(\alpha\beta) = 30$;
3. Ipomoea purpurea: $(AB) = 63$, $(\alpha B) = 18$, $(A\beta) = 10$, $(\alpha\beta) = 55$;

hieraus berechnet sich bei

$$\begin{array}{lll} 1: & \mathfrak{d} = 2{,}500 & q = 0{,}294 \\ 2: & \mathfrak{d} = 7{,}000 & q = 0{,}490 \\ 3: & \mathfrak{d} = 22{,}500 & q = 0{,}901 \end{array}$$

4) Statur von Ehegatten.

Galton[1]) hat bei seinen Untersuchungen über natürliche Vererbung unter anderem der Frage seine Aufmerksamkeit zugewendet, wie sich bei Eheschließungen die Auslese nach der Statur vollzieht. Nachstehend ist das bezügliche Beobachtungsmaterial angegeben.

Tab. 1.

		Gattin		
		groß	mittel	klein
Gatte	groß	18	28	14
	mittel	20	51	28
	klein	12	25	9

[1]) Vgl. G. U. Yule, Philosophical Transactions of the Royal Society of London, A, vol. 194 (1900), p. 292.

Man kann aus dieser neungliedrigen Tabelle zwei viergliedrige konstruieren, indem man auf groß und klein Nachdruck legt, nämlich

		Gattin	
		groß	nicht groß
Gatte	groß	18	42
	nicht groß..	32	113

Aus der ersten Gruppierung ergibt sich der Abhängigkeitskoeffizient

$$q = \frac{690}{3378} = 0{,}204,$$

		Gattin	
		klein	nicht klein
Gatte	klein	9	37
	nicht klein..	42	117

aus der zweiten

$$q = -\frac{501}{2607} = -0{,}192,$$

in so wesentlicher Art hängt also das Ergebnis von der Art der Zusammenfassung des Materials ab. Der große Wechsel von der ersten zur zweiten erklärt sich aus den Zahlen selbst: Man beachte, daß 30% der großen Männer auch große Frauen und 36% der großen Frauen große Männer nehmen; daß hingegen nur 19,6% der kleinen Männer auch kleine Frauen und 17,6% der kleinen Frauen kleine Männer wählen; zwischen großen Personen findet also eine weit stärkere Anziehung statt als zwischen ausgesprochen kleinen. Die stärkste Anziehung zeigen Personen von Mittelgröße; die Zahl ihrer Verbindungen beträgt 24,9% aller.

5) Abhängigkeit von Gebrechen.

Nach der Reichsgebrechlichenzählung 1925/26[1]) gab es im Deutschen Reich unter den 30 196 823 männlichen und 32 213 796 weiblichen Einwohnern 453 495 männliche und 260 076 weibliche Gebrechliche. Es wurden vier Arten von Gebrechen in Betracht gezogen:

A: Blindheit,
B: Taubstummheit oder Taubheit,
C: Körperliche Gebrechen,
D: Geistige Gebrechen.

[1]) Statistisches Jahrbuch für das Deutsche Reich 1932, S. 408.

Die Anzahl der Gebrechlichen wird in der folgenden Übersicht zusammengestellt.

Tab. 2.

Gebrechen	Zahl der Gebrechlichen		Gebrechen	Zahl der Gebrechlichen	
	männlich	weiblich		männlich	weiblich
A	19 157	14 035	AB	261	306
B	23 818	21 558	AC	938	731
C	307 413	122 241	AD	751	655
D	116 514	113 598	BC	780	628
			BD	1505	1428
			CD	9429	7854

Hieraus berechnen sich für die Abhängigkeit von Gebrechen die folgenden q-Werte:

Tab. 3.

Gebrechen	q		Gebrechen	q	
	männlich	weiblich		männlich	weiblich
AB	0,893	0,942	BC	0,535	0,776
AC	0,668	0,871	BD	0,893	0,906
AD	0,828	0,866	CD	0,796	0,908

Aus der vorstehenden Übersicht geht hervor, daß die q-Werte durchgängig positiv sind, d. h. es besteht für sämtliche Gebrechen positive Abhängigkeit. Weiter lehrt die Übersicht, daß die q-Werte für die Frauen ausnahmslos größer sind als für die Männer. Es ist also der Grad der Abhängigkeit zweier Gebrechen beim weiblichen Geschlecht größer als beim männlichen. Ferner ist bemerkenswert, daß q den niedrigsten Wert bei beiden Geschlechtern für die Kombination Taubstummheit und körperliche Gebrechen aufweist. Betrachtet man die aufsteigende Folge der q-Werte, und zwar zunächst für die von D freien Kombinationen, so beobachtet man bei Männern und Frauen die gleiche Folge, nämlich BC, AC, AB. Die Kombinationen mit D schieben sich in diese Reihe ein, und zwar tritt CD bei den Männern als erste (unterste) und bei den Frauen als letzte (oberste) D-Kombination auf. Die beiden anderen D-Kombinationen folgen bei beiden Geschlechtern in gleicher Weise aufeinander. Ob diesen statistischen Beziehungen tiefergehende sachliche Zusammenhänge zugrunde liegen, soll hier nicht weiter untersucht werden. Die Ausnahmestellung von CD wird sicherlich damit zusammenhängen, daß körperliche Gebrechen auf die Frau viel stärker deprimierend wirken als auf den Mann.

§ 3. Mittelbare Abhängigkeit.

16. Vor Anstellung weiterer Betrachtungen sei nochmals als Hauptergebnis folgendes wiederholt. Wenn in einem Kollektiv von N Gliedern zwei Merkmale A, B so verteilt sind, daß die Klasse AB häufiger oder seltener auftritt als bei Unabhängigkeit von A und B zu erwarten wäre, was in den Beziehungen

$$\frac{(AB)}{N} > \frac{(A)}{N}\frac{(B)}{N}, \quad \text{bzw.} \quad \frac{(AB)}{N} < \frac{(A)}{N}\frac{(B)}{N}$$

oder auch

$$(AB) > \frac{(A)(B)}{N}, \quad \text{bzw.} \quad (AB) < \frac{(A)(B)}{N}$$

seinen Ausdruck findet, so spricht man, wenn die Unterschiede erheblich genug sind, um nicht als bloße zufällige Störungen einer wirklich bestehenden Unabhängigkeit gelten zu können, von positiver, bzw. negativer Abhängigkeit.

Wenn man nun auf die Verursachung der Abhängigkeit tiefer eingeht, so kann es zweierlei geben: Entweder betrifft sie die beiden ins Auge gefaßten Merkmale allein oder sie ist auf andere Merkmale zurückführbar, die dabei eine vermittelnde Rolle spielen oder die letzte Ursache der beobachteten Abhängigkeit sind. Ob das eine oder das andere zutrifft, darüber lassen sich zunächst nur Vermutungen aufstellen; aber die statistische Methode bietet auch Mittel und Wege, solche Vermutungen oder Annahmen auf ihre Berechtigung zu prüfen.

Eine Abhängigkeit von der erstgeschilderten Art kann passend als eine unmittelbare, eine der zweiten Art als eine mittelbare bezeichnet werden. Wenn dafür die Benennungen totale und partielle Abhängigkeit gebraucht werden, so hat dies seinen Grund in dem Umstande, daß im ersten Falle auf das ganze Kollektiv Bezug genommen wird, während im andern ein Teilkollektiv die Vergleichsgrundlage bildet.

Als erstes Erläuterungsbeispiel nehmen wir an, in einer Anzahl von Schulen sei eine Statistik der Schüler aufgemacht worden, die sich auf gewisse Gebrechen richtet. Als solche seien behandelt worden: A mangelhafte Körperentwicklung, B Nervenstörungen, C geistige Zurückgebliebenheit. Man kann die Frage aufwerfen, ob nicht B eine vermittelnde Rolle spielt zwischen A und C, d. h. ob nicht die Nervenstörungen auf eine Abnormität im Gehirn hinweisen, die einerseits die körperliche, andererseits die geistige Entwicklung beeinträchtigt hat, so daß im Vorhandensein von B eine Ursache des Zusammentreffens von A und C zu erblicken wäre. Um dies zu prüfen, wird man sich nicht auf das Gesamtkollektiv allein beschränken dürfen, sondern auch entsprechend gewählte Teilkollektive heranziehen müssen.

Ein anderes Beispiel bietet die folgende Materie. Man habe in einer Gesamtheit von Personen, die teils geimpft, teils ungeimpft waren, das Auftreten von Blattern beobachtet und gefunden, daß Nichtgeimpfte beträchtlich häufiger an Blattern erkranken als zu erwarten wäre, wenn Geimpftsein und an Blattern Erkranken unabhängig wären. Dies kann seine Ursache darin haben, daß Impfung einen Schutz gegen Blatternerkrankung gewährt — das wäre eine unmittelbare Abhängigkeit; aber auch eine andere Erklärungsweise bietet sich dar; die Erfahrung lehrt nämlich, daß die Nichtgeimpften zumeist den unteren Volksschichten angehören, die unter unhygienischen Verhältnissen leben, und daß daher das

Zusammentreffen von Nichtgeimpftsein und an Blattern Erkranken vielmehr auf die unhygienischen Verhältnisse zurückzuführen wäre, unter welchen die Nichtgeimpften leben. Wenn also Nichtgeimpftsein, an Blattern Erkranken und das Leben unter gesundheitsschädlichen Verhältnissen mit A, B, C bezeichnet werden, so würde die zwischen A und B aufscheinende Abhängigkeit eine Folge der Abhängigkeit beider von C sein. Um zu entscheiden, ob dem so sei, brauchte man aus dem Gesamtkollektiv nur ein Teilkollektiv Nichtgeimpfter auszulösen, die nicht unter unhygienischen Verhältnissen leben; macht sich unter diesen die Abhängigkeit zwischen A und B auch geltend, dann ist der zweite Erklärungsgrund hinfällig.

Zwischen gewissen Eigenschaften der Großväter und Enkel möge eine Abhängigkeit bemerkbar sein, indem die Eigenschaften bei beiden öfter zusammentreffen als dies nach ihrer Häufigkeit unter den Großvätern einerseits und den Enkeln andererseits zu erwarten wäre, wenn Unabhängigkeit bestünde. Man kann fragen, ist diese Abhängigkeit eine unmittelbare, d. h. findet gewissermaßen ein Überspringen der Eigenschaften von den Großvätern auf die Enkel statt, oder ist sie eine Folge der Übertragung der Eigenschaften der Väter auf die Kinder. Wenn das Vorhandensein der betreffenden Eigenschaft bei Kind, Vater und Großvater mit A, B, C bezeichnet wird, so geht die Frage dahin, ob die Abhängigkeit zwischen A und C nicht hervorgeht aus der Abhängigkeit von A und B und von B und C, wobei also B eine Vermittlungsrolle übernommen hätte. Die Entscheidung über diese Frage könnte durch ein Teilkollektiv herbeigeführt werden, in welchem sämtliche Väter die Eigenschaft entweder besitzen oder nicht besitzen; bliebe auch dann die Abhängigkeit zwischen Großvater und Enkel im früheren Ausmaße bestehen, so wäre sie tatsächlich als eine unmittelbare aufzufassen.

17. Die Beziehungen zur Feststellung solcher mittelbarer Abhängigkeiten ergeben sich aus den früheren Aufstellungen dadurch, daß man an die Stelle des Gesamtkollektivs ein der gestellten Frage entsprechendes Teilkollektiv treten läßt.

Will man erforschen, ob die Abhängigkeit zwischen A und B eine Folge von C ist, so vergleiche man die Häufigkeit der Verbindung ABC mit derjenigen, die sich aus dem Zusammentreffen von AC und BC in dem Kollektiv C, d. h. in der Gesamtheit der Glieder mit der Eigenschaft C, ergeben würde, wenn Unabhängigkeit bestünde; findet man statt der Gleichheit die Ungleichheit

$$(ABC) > \frac{(AC)(BC)}{(C)}, \tag{17}$$

so ist damit eine positive Abhängigkeit zwischen A und B in Bezug auf C nachgewiesen.

Man kann dieser Beziehung eine Reihe anderer Formen geben, die je nach der Art der statistischen Daten Verwendung finden können. So hat man ohne weitere Rechnung

$$\frac{(ABC)}{(AC)} > \frac{(BC)}{(C)} \qquad \frac{(ABC)}{(BC)} > \frac{(AC)}{(C)}. \tag{18}$$

Aus (17) folgt

$$(ABC)(C) > (AC)(BC);$$

ersetzt man darin (C) durch $(BC)+(\beta C)$, (AC) durch $(ABC)+(A\beta C)$, so ergibt sich

$$\frac{(ABC)}{(BC)} > \frac{(A\beta C)}{(\beta C)}, \tag{19}$$

ersetzt man hingegen (C) durch $(AC)+(\alpha C)$ und (BC) durch $(ABC)+(\alpha BC)$, so kommt man zu der Beziehung

$$\frac{(ABC)}{(AC)} > \frac{(\alpha BC)}{(\alpha C)}. \tag{20}$$

Man braucht nur Buchstaben zu vertauschen, wenn von einem andern der drei Merkmale statt von C vermutet wird, daß es die vermittelnde Stellung einnimmt; ist es z. B. B, so treten an die Stelle der vorstehenden Relationen die folgenden:

$$(ABC) > \frac{(AB)(BC)}{(B)} \tag{17*}$$

$$\frac{(ABC)}{(AB)} > \frac{(BC)}{(B)} \qquad \frac{(ABC)}{(BC)} > \frac{(AB)}{(B)} \tag{18*}$$

$$\frac{(ABC)}{(BC)} > \frac{(AB\gamma)}{(B\gamma)} \tag{19*}$$

$$\frac{(ABC)}{(AB)} > \frac{(\alpha BC)}{(\alpha B)}. \tag{20*}$$

18. Beispiele. 1) Erhebungen an 50000 Kindern in England 1892—1894[1]), unter denen sich A körperlich zurückgebliebene, B mit Nervenstörungen behaftete und C geistig schwache Kinder befanden, ergaben folgende auf $N = 10000$ bezogene Zahlen:

$$\begin{array}{lll} (A) = 878 & (B) = 1085 & (C) = 789 \\ (AB) = 337 & (AC) = 338 & (BC) = 455 \\ & (ABC) = 153. & \end{array}$$

Sprechen diese Daten dafür, daß B gemeinsame Ursache von A und C ist, daß also die etwa bemerkbare Abhängigkeit zwischen A und C in B ihren Grund hat?

Daß eine solche Abhängigkeit besteht, zeigt der Vergleich der Häufigkeiten $\frac{(AC)}{(A)}$ und $\frac{(C)}{N}$, einerseits also der geistig Schwachen unter den körperlich Zurückgebliebenen und andererseits der geistig Zurückgebliebenen im Gesamtkollektiv; es ist, in Prozenten ausgedrückt,

$$100\,\frac{(AC)}{(A)} = 38{,}5, \qquad 100\,\frac{(C)}{N} = 7{,}9,$$

[1]) F. Warner, Mental and Physical Conditions among 50000 Children seen 1892—1894, and the Methods of Studying Recorded Observations, with Special Reference to the Determination of the Causes of Mental Dulness and other Defects. Journal of the Royal Statistica Society, vol. 59, part I, 1896, p. 158.

$\frac{(AC)}{(A)} > \frac{(C)}{N}$ in so erheblichem Maße, daß eine starke positive Abhängigkeit außer Zweifel ist.

Im Sinne des ersten Ansatzes in (18*) ergibt sich

$$100\,\frac{(ABC)}{(AB)} = 45{,}4, \qquad 100\,\frac{(BC)}{(B)} = 41{,}9;$$

läßt man β an die Stelle von B treten, so wird

$$100\,\frac{(A\beta C)}{(A\beta)} = 100\,\frac{(AC)-(ABC)}{(A)-(AB)} = 34{,}2, \qquad 100\,\frac{(\beta C)}{(\beta)} = 100\,\frac{(C)-(BC)}{N-(B)} = 3{,}7$$

Wohl ist

$$\frac{(ABC)}{(AB)} > \frac{(BC)}{(B)} \text{ und auch } \frac{(A\beta C)}{(A\beta)} > \frac{(\beta C)}{(\beta)},$$

also positive Abhängigkeit zwischen körperlicher und geistiger Zurückgebliebenheit sowohl bei denjenigen vorhanden, die Nervenstörungen aufweisen, als auch bei denjenigen, die keine Anzeichen davon tragen, aber die erstere ist der letzteren gegenüber von sehr schwachem Grade ($45{,}4 - 41{,}9 = 3{,}5$ gegen $34{,}2 - 3{,}7 = 30{,}5$); das spricht nicht dafür, daß beide Eigenschaften auf Hirndefekte als gemeinsame Ursache zurückzuführen seien; denn sonst müßte das umgekehrte Größenverhältnis bestehen.

2) Aus Galtons[1]) Untersuchungen über die Erblichkeit ist eine Statistik über die Augenfarbe von Großeltern, Eltern und Kindern zu entnehmen, die sich auf 78 kinderreiche Familien zu mindestens 6 Kindern stützt. Wenn Helläugigkeit bei Kind, Eltern und Großeltern (Eltern und Großeltern nach Personen gezählt) mit A, B, C, Dunkeläugigkeit mit α, β, γ bezeichnet wird, so gibt diese Statistik folgende Häufigkeiten:

$$\begin{array}{llll} (ABC) = 1928 & (AB\gamma) = 596 & (A\beta C) = 552 & (A\beta\gamma) = 508 \\ (\alpha BC) = 303 & (\alpha B\gamma) = 225 & (\alpha\beta C) = 395 & (\alpha\beta\gamma) = 501. \end{array}$$

Schon der Anblick dieser Zahlen gibt gewisse Aufschlüsse: lichte Augen durch alle drei Generationen kommen viel häufiger vor als dunkle Augen; am seltensten steht ein lichtäugiger Vater (bzw. eine lichtäugige Mutter) zwischen einem dunkeläugigen Großvater (bzw. einer dunkeläugigen Großmutter) und Kind. Hier handelt es sich um die Frage, ob die zwischen Enkelkindern und Großvätern (bzw. Großmüttern) etwa vorhandene Abhängigkeit eher eine unmittelbare als eine durch den Vater (bzw. die Mutter) vermittelte ist.

Zu diesem Zwecke prüfen wir a) die prozentuale Häufigkeit lichtäugiger Eltern unter licht- und dunkeläugigen Großeltern:

[1]) F. Galton, Natural Inheritance, London 1889, p. 216; vgl. G. U. Yule, An Introduction to the Theory of Statistics, London 1932, p. 46.

$$100\,\frac{(BC)}{(C)} = \frac{223100}{3178} = 70{,}2 \qquad 100\,\frac{(B\gamma)}{(\gamma)} = \frac{82100}{1830} = 44{,}9;$$

b) die prozentuale Häufigkeit lichtäugiger Kinder unter licht- und dunkeläugigen Eltern:

$$100\,\frac{(AB)}{(B)} = \frac{252400}{3052} = 82{,}7 \qquad 100\,\frac{(A\beta)}{(\beta)} = \frac{106000}{1956} = 54{,}2;$$

c) die prozentuale Häufigkeit lichtäugiger Kinder unter licht- und dunkeläugigen Großeltern:

$$100\,\frac{(AC)}{(C)} = \frac{248000}{3178} = 78{,}0 \qquad 100\,\frac{(A\gamma)}{(\gamma)} = \frac{110400}{1830} = 60{,}3.$$

In allen drei Fällen besteht positive Abhängigkeit bezüglich beiderseitiger Lichtäugigkeit, die in den Fällen a) und b), die gleichartig sind, weil sie Eltern und deren Nachkommen betreffen, annähernd von gleichem, im Falle c) von erheblich geringerem Grade ist, wie es die Differenzen

$$70{,}2 - 44{,}9 = 25{,}3, \qquad 82{,}7 - 54{,}2 = 28{,}5, \qquad 78{,}0 - 60{,}3 = 17{,}7$$

anzeigen.

Zur endgültigen Entscheidung der gestellten Frage ist es notwendig, d) die prozentuale Häufigkeit lichtäugiger Enkelkinder unter licht- und dunkeläugigen Großeltern unter Voraussetzung eines lichtäugigen, e) eines dunkeläugigen Zwischengliedes zu berechnen, nämlich:

$$100\,\frac{(ABC)}{(BC)} = \frac{192800}{2231} = 86{,}4 \qquad 100\,\frac{(AB\gamma)}{(B\gamma)} = \frac{59600}{821} = 72{,}6$$

$$100\,\frac{(A\beta C)}{(\beta C)} = \frac{55200}{947} = 58{,}3 \qquad 100\,\frac{(A\beta\gamma)}{(\beta\gamma)} = \frac{50800}{1009} = 50{,}3.$$

Es besteht auch unter diesem Gesichtspunkte eine, nunmehr mittelbare, Abhängigkeit, ebenfalls positiv, aber beträchtlich schwächer; denn die Differenzen

$$86{,}4 - 72{,}6 = 13{,}8. \qquad 58{,}3 - 50{,}3 = 8{,}0$$

sind erheblich geringer als bei der unmittelbaren Abhängigkeit, wo die Differenz mindestens 17,7 betrug. Es kann also die zwischen Enkelkindern und Großeltern bestehende Abhängigkeit nicht auf die Übertragung durch die Väter allein zurückgeführt werden, vielmehr ist eine Komponente unmittelbarer Abhängigkeit anzunehmen.

3) Die Volkszählung im Deutschen Reich (ohne Württemberg und Baden, weil hier die Altersgliederung nicht vorliegt) vom Jahre 1925 hat folgende Statistik über die Verbreitung A der Blindheit, B der Taubstummheit oder Taubheit, D der Geistesgestörtheit und von Verbindungen dieser Gebrechen geliefert:

Tab. 4.

Bezeichnung der Gesamtheit	I. Alle männlichen Personen[1])	II. Männliche Personen im Alter von 60 Jahren und darüber[2])
N	27 837 839	2 415 054
(A)	17 395	5 764
(B)	21 185	1 843
(D)	105 374	11 821
(AB)	234	105
(AD)	685	128
(BD)	1 246	132
(ABD)	31	11

Es ist die Häufigkeit der einzelnen Gebrechen bei den Kategorien I und II, sodann die unmittelbare Abhängigkeit zwischen Blindheit und Geistesgestörtheit, endlich die mittelbare Abhängigkeit dieser beiden Gebrechen unter den Taubstummen zu untersuchen; sämtliche Häufigkeiten sind auf 1000 zu beziehen.

Zur Erledigung der ersten Frage sind die Tausendfachen der Quotienten

$$\frac{(A)}{N} \qquad \frac{(B)}{N} \qquad \frac{(D)}{N}$$

zu bilden; bei der zweiten Frage kommt es auf die Vergleichung der gleichfalls mit 1000 vervielfachten Quotientenpaare

$$\frac{(AD)}{(D)}, \quad \frac{(A\delta)}{(\delta)} \quad \text{bei I und II}$$

an; für die dritte Frage sind die Quotienten

$$\frac{(ABD)}{(BD)} \quad \frac{(AB\delta)}{(B\delta)} \quad \text{für I und II}$$

in Promille zu bilden.

Das Ergebnis dieser Rechnungen ist in der umstehenden Tab. 5 zusammengestellt.

Aus dieser Gegenüberstellung ist folgendes zu entnehmen:

Unter den drei Gebrechen ist Geistesgestörtheit (bei I und II) am stärksten, Blindheit (bei I), bzw. Taubstummheit (bei II) am schwächsten vertreten. Blindheit und Geistesgestörtheit kommen in den hohen Altern häufiger vor als in der

[1]) Statistisches Jahrbuch für das Deutsche Reich 1932, S. 5 u. 408.

[2]) Statistik des Deutschen Reichs, Band 401, I, 1928, S. 236 u. f., und Band 419, 1931, S. 11, 26, 57 u. 97.

Tab. 5.

Art des Gebrechens	I. Häufigkeit bei allen männlichen Personen pro 1000	II. Häufigkeit bei männlichen Personen im Alter von 60 Jahren und darüber pro 1000
Blindheit	0,62	2,39
Taubstummheit	0,76	0,76
Geistesgestörtheit	3,79	4,89
Blindheit unter Geistesgestörten	6,50	10,83
Blindheit unter geistig Normalen	0,60	2,35
Blindheit unter geistesgestörten Taubstummen	24,88	83,33
Blindheit unter geistig normalen Taubstummen	10,18	54,94

Gesamtheit, die Taubstummheit ist unter den Bejahrten ebenso häufig vertreten wie in der Gesamtheit.

Zwischen Blindheit und Geistesgestörtheit herrscht (bei I und II) positive Abhängigkeit. Beide Gebrechen zusammen treten im Alter in verstärktem Maße auf.

Zwischen Blindheit und Geistesgestörtheit unter den Taubstummen besteht ebenfalls, und zwar in hohem Maße positive Abhängigkeit; die beiden erstgenannten Gebrechen sind unter den Taubstummen hohen Alters viel stärker verbreitet als in der Gesamtbevölkerung (83,33 gegen 24,88), aber auch die Blindheit allein ist unter ihnen viel häufiger als in der ganzen Bevölkerung (54,94 gegen 10,18).

§ 4. Mehrfache Klassifikation.

19. Die bisherigen Betrachtungen hatten es mit Merkmalen und Verbindungen solcher zu tun, deren jedes nur zwei Alternativen zuläßt: sein Vorkommen und sein Fehlen oder das Vorkommen des Gegensatzes. Dies ist die einfachste Gliederung eines auf Zählung beruhenden statistischen Materials.

Der allgemeine, auf den verschiedensten Gebieten wiederkehrende Fall besteht darin, daß mehrere Merkmale in Betracht gezogen werden, bei deren jedem eine bestimmte Anzahl von Arten, Stufen o. dgl. zu unterscheiden ist. Sind A, B, ... K die Merkmale,

$$\begin{array}{l} A_1, \ A_2, \ \ldots\ldots A_i, \ \ldots A_m \\ B_1, \ B_2, \ \ldots\ldots B_j \ \ldots B_n \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ K_1, \ K_2, \ \ldots\ldots K_p, \ \ldots K_s \end{array}$$

ihre Abarten, so sind die Klassen durch alle möglichen Verbindungen von der Zusammensetzung

$$A_i \ B_j \ \ldots\ldots K_p$$

gebildet; die Häufigkeitszahl werde mit $(A_i\ B_j\ \ldots . K_p)$ bezeichnet, die Anzahl solcher Klassen der höchsten Ordnung beträgt $mn \ldots s$.

Bei einer größeren Anzahl der Merkmale und ihrer Abarten fällt es außerordentlich schwer, einen Einblick in die große Mannigfaltigkeit möglicher Beziehungen zu gewinnen.

Eine leichtere Übersicht gestattet noch der Fall zweier Merkmale A, B, die sich in m, bzw. n Abarten spalten. Da gerade dieser Fall sehr häufig vorkommt, so ist es geboten, näher auf ihn einzugehen; man kann die auf ihn bezüglichen Betrachtungen und Untersuchungen als Theorie der Tafeln mit doppeltem Eingang bezeichnen.

In der Tat, wenn $A_1, A_2, \ldots A_i, \ldots A_m$ zur Bezeichnung der Kolonnen, $B_1, B_2, \ldots B_j, \ldots B_n$ zur Bezeichnung der Zeilen verwendet werden, so entsteht eine Tafel mit mn Feldern, deren jedes einer bestimmten der mn Klassen des Kollektivs zugeordnet ist; das Feld, in welchem die $i$te Kolonne die $j$te Zeile kreuzt, enthält die Zahl $(A_i\ B_j)$.

Die Kolonnensummen

$$\sum_1^n {}^j (A_i\ B_j) = (A_i)$$

geben die Verteilung des Merkmals A ohne Rücksicht auf B, die Zeilensummen

$$\sum_1^m {}^i (A_i\ B_j) = (B_j)$$

die Verteilung des Merkmals B ohne Rücksicht auf A; endlich ist

$$\sum_1^m (A_i) = \sum_1^n (B_j) = N$$

der Gesamtumfang des Kollektivs. Das Gerippe einer solchen Tafel bietet also das folgende Bild dar:

	A_1	A_2	A_3	.	.	.	
B_1	$(A_1 B_1)$	$(A_2 B_1)$	$(A_3 B_1)$	.	.	.	(B_1)
B_2	$(A_1 B_2)$	$(A_2 B_2)$	$(A_3 B_2)$	.	.	.	(B_2)
B_3	$(A_1 B_3)$	$(A_2 B_3)$	$(A_3 B_3)$	.	.	.	(B_3)
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
	(A_1)	(A_2)	(A_3)	.	.	.	N

Um eine bessere Übersicht über die Verteilung der Merkmale zu erhalten und verschiedene Materien derselben Art in dieser Hinsicht miteinander vergleichen zu können, empfiehlt es sich, N auf einen festen Wert, 1, 100, 1000 oder eine höhere Potenz von 10 zurückzuführen und auf Grund der vorstehenden Tafel eine neue zu berechnen mit den Werten

$$\frac{(A_i\ B_j)}{N} \text{ oder } 100\ \frac{(A_i\ B_j)}{N} \text{ oder } 1000\ \frac{(A_i\ B_j)}{N}.$$

Aus einer solchen Tafel können die Antworten auf die mannigfachsten Fragen durch entsprechende Zusammenfassung und Inbeziehungsetzung der Zahlen gewonnen werden.

Nehmen wir beispielsweise an, A bedeute die Todesursache, B den Beruf eines Verstorbenen. $A_1, A_2, \ldots A_m$ sind dann die verschiedenen Todesursachen, die man unterscheiden will, $B_1, B_2, \ldots B_n$ die verschiedenen Berufe, die in Betracht kommen. Man kann dann mehrere Krankheitsformen aus irgend einem Gesichtspunkte zusammenfassen und darnach fragen, in welchem Maße eine Gruppe von Berufen an diesen Todesursachen beteiligt ist, aber auch darnach, welchen Anteil mehrere Todesursachen an den Todesfällen bestimmter Berufe haben. Man wird zu diesem Zwecke in der ursprünglichen Tabelle die Felder, die den ausgewählten Todesursachen und den ausgewählten Berufen gemeinsam sind, addieren und diese Summe im ersten Falle zur Summe der betreffenden (B_j), im andern Falle zur Summe der betreffenden (A_i) ins Verhältnis setzen. So ist z. B.

$$100 \frac{(A_1 B_1) + (A_2 B_1) + (A_3 B_1) + (A_1 B_2) + (A_2 B_2) + (A_3 B_2)}{(B_1) + (B_2)}$$

der Prozentsatz, mit welchem die Berufe B_1, B_2 zusammen an den Todesursachen A_1, A_2, A_3 beteiligt sind, und

$$100 \frac{(A_1 B_1) + (A_2 B_1) + (A_3 B_1) + (A_1 B_2) + (A_2 B_2) + (A_3 B_2)}{(A_1) + (A_2) + (A_3)}$$

der Prozentsatz, mit welchem die Todesursachen A_1, A_2, A_3 zusammen an den Sterbefällen innerhalb der Berufe B_1, B_2 teilnehmen.

20. Wenn es sich weiter darum handelt, die Abhängigkeiten zwischen den Unterarten von A und B zu erforschen, so wird diese Arbeit aufzulösen sein in die Untersuchung von Quadrupeln, d. i. von vier Feldern, die zu je zwei in einer Zeile und Kolonne liegen, wofür die Ausführungen von Artikel 14 maßgebend sind.

Dazu ist die Ausrechnung derjenigen Verteilung erforderlich, die platzgriffe, wenn Unabhängigkeit bestünde. Zu jedem $(A_i B_j)$ gehört ein Unabhängigkeitswert

$$[A_i B_j] = \frac{(A_i)(B_j)}{N}. \tag{21}$$

Die Zusammenstellung dieser Werte gibt eine der früheren konforme Tabelle, die mit ihr in den vier Rändern vollständig übereinstimmt:

	A_1	A_2	A_3	.	.	.	
B_1	$[A_1 B_1]$	$[A_2 B_1]$	$[A_3 B_1]$	.	.	.	(B_1)
B_2	$[A_1 B_2]$	$[A_2 B_2]$	$[A_3 B_2]$	.	.	.	(B_2)
B_3	$[A_1 B_3]$	$[A_2 B_3]$	$[A_3 B_3]$	.	.	.	(B_3)
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
	(A_1)	(A_2)	(A_3)	.	.	.	N

Schon der bloße vergleichende Blick auf die beiden Tabellen gestattet ein allgemeines Urteil darüber, ob eine ausgesprochene Abhängigkeit besteht oder nicht. Bei vollkommener Unabhängigkeit, die auch von zufälligen Störungen frei sein müßte, würden beide Tabellen vollständig übereinstimmen.

Vor allem sei festgestellt, daß die Determinante aus jedem Quadrupel, das dieser Tabelle entnommen wird, den Wert Null hat; es genügt, dies an

$$\begin{vmatrix} [A_1 B_1] & [A_2 B_1] \\ [A_1 B_2] & [A_2 B_2] \end{vmatrix}$$

zu zeigen; nach (21) gibt diese Determinante entwickelt und mit Fortlassung des Nenners N^2

$$(A_1)(B_1)(A_2)(B_2) - (A_1)(B_2)(A_2)(B_1).$$

Hat die entsprechende Determinante aus der Tabelle der beobachteten Häufigkeiten

$$\begin{vmatrix} (A_1 B_1) & (A_2 B_1) \\ (A_1 B_2) & (A_2 B_2) \end{vmatrix}$$

einen positiven (negativen) Wert, so besteht zwischen A und B innerhalb des Bereichs, der durch die Indizes angezeigt ist, positive (negative) Abhängigkeit in um so stärkerem Maße, je größer der absolute Betrag der Determinante ist. Man kann die Bedingung der positiven Abhängigkeit auch in eine der Formen

$$\frac{(A_1 B_1)}{(A_2 B_1)} > \frac{(A_1 B_2)}{(A_2 B_2)} \tag{22}$$

$$\frac{(A_1 B_1)}{(A_1 B_2)} > \frac{(A_2 B_1)}{(A_2 B_2)} \tag{23}$$

kleiden, bei negativer Abhängigkeit tritt das Zeichen $<$ in Kraft. Bildet man also nach (22) die Quotienten der in der Tabelle nebeneinander, nach (23) die Quotienten untereinander stehender Klassenhäufigkeiten, so zeigt ihr Anwachsen positive, ihre Abnahme negative Abhängigkeit an.

Hält in einer Tafel das Ansteigen oder das Abfallen benachbarter Quotienten an, so besteht es, wie unmittelbar einzusehen, auch zwischen nicht benachbarten Quotienten; eine so beschaffene Tafel bezeichnet man als isotrop. Freilich hängt diese Eigenschaft von der Ordnung der A und der B ab, kann durch deren Änderung zerstört, aber auch herbeigeführt werden; die Regel wird Anisotropie bilden.

Auf diese Weise kann man alle Abhängigkeitsfragen zur Lösung bringen und über alle Einzelheiten der Materie Aufschluß erlangen.

21. Sowie es üblich ist, die Genauigkeit einer Beobachtungsreihe durch eine einzige Zahl, etwa den mittleren oder den wahrscheinlichen Fehler zu kennzeichnen, so kann auch das Verlangen sich einstellen, die Abhängigkeitsverhältnisse innerhalb einer Materie von der vorliegenden Art durch eine Zahl zum Ausdruck zu bringen. Eine solche Zahl hat K. Pearson[1]) in Vorschlag gebracht, ihre Konstruktion beruht auf folgenden Erwägungen.

[1]) Drapers' Company Research Memoirs, Biometric Series 1, 1904.

Bei vollständiger Unabhängigkeit ist jedes

$$(A_i B_j) = [A_i B_j],$$

daher jedes

$$(A_i B_j) - [A_i B_j] = 0;$$

die Regel aber ist, daß jedes

$$(A_i B_j) - [A_i B_j] = \delta_{ij} \tag{24}$$

einen von Null verschiedenen, bald positiven, bald negativen Wert hat.

Auf diese δ_{ij} muß sich die zu bildende Zahl stützen; doch ist zweierlei zu beachten: wenn nicht Isotropie besteht, wird wegen des wechselnden Zeichens vom Vorzeichen abgesehen werden müssen, ehe man an eine Vereinigung dieser Differenzen schreitet, was am einfachsten durch Quadrieren erreicht wird; derselbe absolute Wert ferner hat bei δ_{ij} nicht die gleiche Bedeutung, wenn es sich um große und um kleine Wiederholungszahlen handelt; je größer die Wiederholungszahl, desto mehr sinkt die Bedeutung einer bestimmten Größe von δ_{ij}. Diesen Rücksichten trägt der Quotient

$$\frac{\delta_{ij}^2}{[A_i B_j]}$$

Rechnung. Pearson verwendet die Summe dieser Quotienten

$$\chi^2 = \Sigma \frac{\delta_{ij}^2}{[A_i B_j]} \tag{25}$$

zur Bildung der Zahl

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}, \tag{26}$$

die er als Gesamtmaß der in der Materie herrschenden Abhängigkeit unter dem Namen Zufälligkeitskoeffizient einführt. Er wird Null bei Unabhängigkeit, weil dann jedes δ_{ij} und somit auch χ^2 Null wird, und er strebt der Einheit zu, wenn mit den δ_{ij}^2 auch χ^2 beständig wächst.

Ersetzt man in (25) δ_{ij} durch seinen Ausdruck und beachtet man weiter, daß die Summe aller $[A_i B_j]$ ebenso wie die Summe aller $(A_i B_j)$ gleich N ist, so ergibt sich

$$\chi^2 = \Sigma \frac{(A_i B_j)^2}{[A_i B_j]} - N,$$

und wird die neu auftretende Summe mit S bezeichnet, so kommt

$$C = \sqrt{\frac{S - N}{S}} \tag{27}$$

Bei dieser zweiten Darstellung entfällt die Bildung der δ_{ik}, hingegen hat man bei der ersten kleinere Zahlen zu quadrieren als bei der zweiten. Indessen wird man umfangreichere Rechnungen dieser Art wohl nicht ohne die Benützung von Quadrattafeln ausführen und dann verdient die Formel (27) den Vorzug.

Die Summe S, die man auch schreiben kann

$$\Sigma (A_i B_j) \frac{(A_i B_j)}{[A_i B_j]}$$

verwandelt sich bei Unabhängigkeit, weil dann jedes $\frac{(A_i B_j)}{[A_i B_j]}$ gleich 1 ist, in $\Sigma (A_i B_j) = N$, womit wieder dargetan ist, daß bei Unabhängigkeit $C = 0$ ist. Die obere Grenze, die C erreichen kann, nämlich 1, würde sich, theoretisch gesprochen, bei $S \to \infty$ einstellen. Sehen wir zu, welchen Wert C bei einer größtmöglichen Abhängigkeit erreichen kann. Von einer solchen wird gewiß gesprochen werden dürfen, wenn bei einer Tafel mit mm Feldern nur die Diagonalfelder Häufigkeitszahlen enthalten, während alle übrigen Felder mit Nullen besetzt sind, wie dies die folgende schematische Tafel für $m = 6$ darstellt.

	A_1	A_2	A_3	A_4	A_5	A_6	
B_1	$(A_1 B_1)$	.	.	.	.	.	(B_1)
B_2	.	$(A_2 B_2)$	.	.	.	.	(B_2)
B_3	.	.	$(A_3 B_3)$	.	.	.	(B_3)
B_4	.	.	.	$(A_4 B_4)$	.	.	(B_4)
B_5	.	.	.	.	$(A_5 B_5)$	.	(B_5)
B_6	.	.	.	.	.	$(A_6 B_6)$	(B_6)
	(A_1)	(A_2)	(A_3)	(A_4)	(A_5)	(A_6)	N

Denn auf das Beispiel mit den Todesursachen und Berufen angewendet, hieße das, daß jede Todesursache nur einen Beruf trifft, also gewiß eine extremale Abhängigkeit vorhanden wäre.

In diesem Falle besteht S nur aus Gliedern der Form $\frac{(A_i B_i)^2}{[A_i B_i]}$, und da bei jedem i die Beziehungen $(A_i B_i) = (A_i) = (B_i)$ und $[A_i B_i] = \frac{(A_i)(B_i)}{N}$ gelten, so reduziert sich jedes Glied der Summe auf N und die Summe selbst auf $m N$, mithin wird

$$C = \sqrt{\frac{m-1}{m}}$$

Aus diesem besondern Falle zieht man den Schluß, daß man den Idealfall, bei großer Abhängigkeit mit C der Einheit so viel wie möglich nahe zu kommen, um so besser erreicht, je größer m, je feiner die Gliederung des Materials. Auf grobe Gliederung ist also der Zufälligkeitskoeffizient nicht gut anwendbar, als zu wenig empfindlich. Bei $m = 5$ kann er die Grenze $\sqrt{\frac{4}{5}} = 0{,}8944$ erreichen.

Als einen Nachteil des Zufälligkeitskoeffizienten wird man seine rein theoretische Konstruktion und darum seine schwere Erfaßbarkeit bezeichnen müssen.

22. Beispiele. 1) Haar- und Augenfarbe.

Die folgende Tabelle gibt das Ergebnis der Erhebungen an 6800 männlichen Personen der badischen Bevölkerung[1]). Es sind dabei vier Haarfarben: A_1 blond, A_2 braun, A_3 schwarz, A_4 rot, und drei Augenfarben: B_1 blau, B_2 grau oder grün, B_3 braun, im ganzen also 12 Klassen $A_i B_j$ unterschieden worden.

Tab. 6. Haar- und Augenfarbe der männlichen badischen Bevölkerung.

Tabelle der $(A_i B_j)$.

Farbe der Augen	Farbe der Haare				
	A_1 blond	A_2 braun	A_3 schwarz	A_4 rot	
B_1 blau	1768	807	189	47	2811 (B_1)
B_2 grau oder grün	946	1387	746	53	3132 (B_2)
B_3 braun	115	438	288	16	857 (B_3)
	(A_1) 2829	(A_2) 2632	(A_3) 1223	(A_4) 116	$6800 = N$

Aus der Tabelle geht hervor, daß es sich um eine vorherrschend blondhaarige, und grau- oder grünäugige Bevölkerung handelt. Rothaarigkeit kommt nur sporadisch vor, häufiger ist sie mit braunen als mit lichten (B_1 und B_2) Augen einhergehend, sie macht dort

$$100 \frac{(A_4 B_3)}{(B_3)} = \frac{1600}{857} = 1{,}9\,\%,$$

hier aber nur

$$100 \frac{(A_4 B_1) + (A_4 B_2)}{(B_1) + (B_2)} = \frac{10000}{5943} = 1{,}7\,\%$$

aus. Schwarze Haare sind an den drei Augenfarben in folgenden prozentualen Verhältnissen beteiligt:

$$100 \frac{(A_3 B_1)}{(B_1)} = \frac{18900}{2811} = 6{,}7\,\%$$

$$100 \frac{(A_3 B_2)}{(B_2)} = \frac{74600}{3132} = 23{,}8\,\%$$

$$100 \frac{(A_3 B_3)}{(B_3)} = \frac{28800}{857} = 33{,}6\,\%,$$

kommen also bei Blauäugigen sehr selten, am häufigsten bei Braunäugigen vor.

In dieser Weise kann man sich über die verschiedensten Einzelfragen Aufschluß verschaffen.

Die Untersuchung der Abhängigkeiten leiten wir mit der Ableitung der Unabhängigkeitshäufigkeiten ein.

[1]) O. Ammon, Zur Anthropologie der Badener. Jena 1899, II. Teil, S. 157.

Tab. 7. Haar- und Augenfarbe der männlichen badischen Bevölkerung.

Tabelle der $[A_i B_j]$.

Farbe der Augen	Farbe der Haare				
	blond	braun	schwarz	rot	
blau	1169,46	1088,02	505,57	47,95	2811
grau oder grün	1303,00	1212,27	563,30	53,43	3132
braun	356,54	331,71	154,13	14,62	857
	2829	2632	1223	116	6800

Die Dezimalstellen sind nur zu dem Zwecke entwickelt worden, um die arithmetischen Eigenschaften dieser Tafel mit voller Schärfe hervortreten zu lassen. Die Abweichungen der beobachteten von den berechneten Zufallshäufigkeiten sind beträchtlich, nur die Rothaarigkeit ist wie durch den Zufall auf die verschiedenen Augenfarben verteilt. Es ist ein charakteristisches Merkmal dieser Tabelle, daß die Determinante jedes Quadrupels Null ist (mit der durch die eingehaltene Schärfe der Rechnung bedingten Annäherung), z. B.

$$\begin{vmatrix} 1088{,}02 & 47{,}95 \\ 331{,}71 & 14{,}62 \end{vmatrix} = 15906{,}9 - 15905{,}5.$$

Die Differenzen $(A_i B_j) - [A_i B_j]$ weisen folgende Vorzeichen auf:

$$\begin{matrix} + & - & - & - \\ - & + & + & - \\ - & + & + & + \end{matrix}$$

woraus zu ersehen ist, bei welchen Farbenkombinationen positive, bei welchen negative Abhängigkeit besteht; so zeigt sich zwischen blonden Haaren und blauen Augen eine beträchtliche Anziehung, zwischen blonden Haaren und grauen oder grünen und braunen Augen eine ausgesprochene Abstoßung.

Nach Vorschrift von (22) und (23) ergeben sich folgende Quotienten:

(22)				(23)			
2,19	4,27	4,02		1,87	0,58	0,25	0,89
0,68	1,86	14,08		8,23	3,17	2,59	3,31
0,26	1,52	18,00					

Die vertikal gebildeten Differenzen in (22) wie die horizontal gebildeten in (23) zeigen folgende Zeichenstellung:

$$\begin{matrix} + & + & - \\ + & + & - \end{matrix}$$

Wie ein Blick auf (23) lehrt, hebt die Versetzung der letzten Kolonne hinter die erste den Zeichenwechsel auf und führt zu einer ständigen Abnahme in

horizontalem Sinne: die Tabelle wird also, wenn man die Haarfarben zu blond, rot, braun, schwarz umordnet, zu einer isotropen.

Im Anschluß hieran soll der Zufälligkeitskoeffizient zur Illustrierung der Methode bestimmt werden.

$\frac{(A_1 B_1)^2}{[A_1 B_1]}$	2672,9	$\frac{(A_1 B_2)^2}{[A_1 B_2]}$	686,8	$\frac{(A_1 B_3)^2}{[A_1 B_3]}$	37,1
$\frac{(A_2 B_1)^2}{[A_2 B_1]}$	598,6	$\frac{(A_2 B_2)^2}{[A_2 B_2]}$	1586,9	$\frac{(A_2 B_3)^2}{[A_2 B_3]}$	578,3
$\frac{(A_3 B_1)^2}{[A_3 B_1]}$	70,7	$\frac{(A_3 B_2)^2}{[A_3 B_2]}$	988,0	$\frac{(A_3 B_3)^2}{[A_3 B_3]}$	538,1
$\frac{(A_4 B_1)^2}{[A_4 B_1]}$	46,1	$\frac{(A_4 B_2)^2}{[A_4 B_2]}$	52,6	$\frac{(A_4 B_3)^2}{[A_4 B_3]}$	17,5
	3388,3		3314,3		1171,0

$$S = 3388,3 + 3314,3 + 1171,0 = 7873,6;$$

hiernach ist

$$C = \sqrt{\frac{1073,6}{7873,6}} = 0,37.$$

Der Erkenntniswert des Zufälligkeitskoeffizienten C kommt bei Vergleichungen zur Geltung, wie das folgende Beispiel zeigt.

2) Ähnlichkeit von Brüdern in athletischen Eigenschaften und von Schwestern im Temperament.

Die Angaben in den folgenden Tafeln stammen von K. Pearson[1]). In beiden bedeuten diesmal A und B ein und dasselbe Merkmal, jedoch an zwei verschiedenen Personen.

Tab. 8a. Verteilung athletischer Eigenschaften auf Brüder.

		Erster Bruder			
		athletisch	mittel	nichtathletisch	
zweiter Bruder	athletisch	906	20	140	1066
	mittel	20	76	9	105
	nichtathletisch	140	9	370	519
		1066	105	519	1690

[1]) K. Pearson, On the Laws of Inheritance in Man. Biometrika, vol. III, London 1904, p. 182, 188.

Tab. 8b. Verteilung der Temperamente unter Schwestern.

		Erste Schwester			
		hitzig	gutmütig	mürrisch	
zweite Schwester	hitzig	198	177	77	452
	gutmütig	177	996	165	1338
	mürrisch	77	165	120	362
		452	1338	362	2152

In Tab. 8a überwiegen die Brüderpaare mit athletischem Bau, und die aus ihnen zusammengesetzte Klasse $A_1 B_1$ ist die zahlreichste.

In Tab. 8b überwiegen die Gutmütigen, und die Klasse $A_2 B_2$, in der beide Schwestern diese Gemütsart aufweisen, ist am stärksten besetzt.

Bei vorausgesetzter Unabhängigkeit würde sich die Verteilung wie folgt gestalten.

Tab. 9. Körperbau.

		Erster Bruder			
		athletisch	mittel	nichtathletisch	
zweiter Bruder	athletisch	672,4 (+ 233,6)	66,2 (− 46,2)	327,4 (− 187,4)	1066
	mittel	66,2 (− 46,2)	6,5 (+ 69,5)	32,2 (− 23,2)	105
	nichtathletisch	327,4 (− 187,4)	32,2 (− 23,2)	159,4 (+ 210,6)	519
		1066	105	519	1690

Tab. 10. Gemütsart.

		Erste Schwester			
		hitzig	gutmütig	mürrisch	
zweite Schwester	hitzig	94,9 (+ 103,1)	281,0 (− 104,0)	76,0 (+ 1,0)	452
	gutmütig	281,0 (− 104,0)	831,9 (+ 164,1)	225,1 (− 60,1)	1338
	mürrisch	76,0 (+ 1,0)	225,1 (− 60,1)	60,9 (+ 59,1)	362
		452	1338	362	2152

Die in Klammern beigesetzten Zahlen sind die Abweichungen δ_{ij}, die in ihrem Vorzeichen erkennen lassen, ob die Abhängigkeit der betreffenden Merkmalvereinigung positiv oder negativ ist. In beiden Fällen weist die Gleichartigkeit der Geschwister positive Abhängigkeit auf, ihre Ungleichartigkeit mit einer belanglosen Ausnahme negative Abhängigkeit.

Zur Bestimmung des C können diese Abweichungen, nachdem sie einmal berechnet sind, verwendet werden gemäß den Formeln (25), (26), man kann aber auch in der andern Art bloß mit den Zahlen $(A_i\ B_j)$, $[A_i\ B_j]$ rechnen und findet auf beide Arten übereinstimmend

$$\text{aus der ersten Tafel } C = 0{,}68,$$
$$\text{aus der zweiten Tafel } C = 0{,}36,$$

wonach die Abhängigkeit in der körperlichen Beschaffenheit der Brüder wesentlich größer sich herausstellt als die Abhängigkeit in der Gemütsart der Schwestern.

Zweiter Abschnitt.

Theorie der veränderlichen Merkmale.

§ 1. Die Verteilungen in Kollektiven.

23. Die Kollektive, die von nun ab den Gegenstand der Betrachtung bilden werden, sind von solcher Art, daß an jedem Gliede ein veränderliches Merkmal (später auch deren zwei oder mehr) in einem bestimmten Grade verwirklicht ist. Wir denken dabei vorerst an den direkten Fall eines quantitativen Merkmals und stellen uns vor, sein Grad, allgemein mit X bezeichnet, sei an jedem Gliede durch Messung oder Zählung bereits festgestellt. Dann gehört zu jedem Gliede des Kollektivs ein besonderer Wert x von X, und die Gesamtheit dieser Zahlwerte ist es, die den Gegenstand der Untersuchung zu bilden hat.

Die Variable X heißt das Argument oder die Ordnungsgröße des Kollektivs, ihre besonderen Werte bilden eine Kollektivreihe, sie ist das arithmetische Abbild des Kollektivs und ihr Umfang stimmt mit dem Umfang des Kollektivs überein.

In Bezug auf die arithmetische Natur von X sind zwei Fälle zu unterscheiden: Entweder ist X eine stetige Variable, die an sich alle Werte innerhalb gewisser Schranken annehmen kann, z. B. eine Dimension, ein Inhalt, ein Gewicht, eine Zeitdauer, die Intensität einer Erscheinung, oder eine unstetige Variable, die nur bestimmte diskrete Werte annehmen kann, z. B. die veränderliche Anzahl gewisser Organe eines Lebewesens. Diese Unterscheidung ist jedoch mehr von theoretischer als von praktischer Bedeutung; wir sind ja nicht imstande, an einer stetigen Größe alle Werte zu konstatieren, die sie annehmen kann, weil die Unvollkommenheit unserer Messungen uns dazu zwingt, daß wir uns mit mehr oder weniger groben Abstufungen begnügen. Dem wissenschaftlichen Bedürfnis kann damit trotzdem gedient sein. Es braucht daher zwischen der Behandlung stetiger und unstetiger Kollektive kein tiefgehender Unterschied gemacht zu werden.

24. Die Aufnahme eines Kollektivs liefert eine Zahlenreihe, seine Urliste. Aus dieser ist noch nicht zu erkennen, wie die beobachteten Werte des Merkmals auf die einzelnen Glieder verteilt sind. Erst nachdem sie arithmetisch geordnet worden ist, läßt sich manches erkennen, so vor allem der kleinste und der größte in dem Kollektiv vorkommende Wert des Arguments; diese begrenzen das Variationsintervall, seine Ausdehnung heiße die Variationsbreite oder -weite.

Durch das Ordnen geht aus der Urliste die primäre Verteilungstafel hervor.

Der nächste Schritt ist die Einteilung des Kollektivs in Klassen. Man legt einen Maßstab für X zugrunde, bestehend in einer Reihe gleich weit voneinander abstehender Werte $X_1, X_2, \ldots X_n$, die zum mindesten das Variationsintervall umspannen, wenn nötig, auch darüber hinaus reichen. Der Abstand benachbarter Werte bildet das Klassenintervall oder die Klassengröße.

Alle Glieder des Kollektivs, deren Argument in ein und dasselbe Klassenintervall fällt, bilden zusammen eine Klasse; ihre Anzahl macht die Klassenhäufigkeit aus. Die Summe der Klassenhäufigkeiten muß mit dem Umfang des Kollektivs übereinstimmen.

Erst die Zusammenfassung der Klassen mit ihren Klassenhäufigkeiten gibt ein taugliches Mittel zur Beschreibung und Untersuchung des Kollektivs und wird seine Verteilungstafel genannt.

Bei der Schaffung dieser Grundlage für alles Weitere können die folgenden Bemerkungen von Wert sein:

a) Für die Wahl des Klassenintervalls ist eine Reihe von Umständen maßgebend: der Umfang des Kollektivs, die Variationsweite, die Maßeinheit, der Zweck der Untersuchung.

Zum Umfang muß die Klassengröße in solcher Beziehung stehen, daß wenigstens im Kern der Tafel alle Klassen besetzt sind. Gegen die Enden zu dürfen leere Klassen vorkommen.

Man wird in der Regel die zugrunde liegende Maßeinheit selbst oder ein Vielfaches derselben zur Klassengröße machen. Ob das erstere und welches Vielfache im andern Falle zu nehmen sei, richtet sich nach der Variationsweite; denn von beiden Umständen hängt die Zahl der Klassen ab. Eine Tafel mit sehr vielen Klassen, etwa 20 und darüber, wird unübersichtlich und beschwerlich für die weitere Behandlung.

Nach getroffener Wahl des Klassengerippes kann es sich als zweckmäßig oder als erforderlich erweisen, zu einer feineren oder gröberen Klassifikation überzugehen. Dabei hat auch der verfolgte Zweck mitzusprechen, der das eine Mal eine größere Schärfe der Resultate verlangt, das andere Mal einen großen Aufwand an Arbeit nicht rechtfertigen würde. Allgemeine Weisungen hierüber lassen sich nicht geben, nur der Weg der Erfahrung kann zu dem Rechten führen.

Bei unstetigen Kollektiven, wo das Argument nur ganzzahlige Werte annehmen kann, werden eben diese Werte auch die Träger der Klassen sein: Der Begriff des Klassenintervalls fällt also hinweg. Er wird indessen künstlich eingeführt, indem man die Klassengrenzen — auch Wechselpunkte — in die Mitte zwischen die Klassenpunkte verlegt.

b) Mit der Lage des Klassengerippes hängt die Bezifferung der Wechselpunkte und der Klassenmitten zusammen. Man wird nicht gern davon abgehen, daß wenigstens die eine Gattung dieser Punkte mit ganzen Zahlen beziffert sei. Wenn z. B. die Klassengröße 1 cm ist, so können entweder die Wechselpunkte ganze Zentimeter bedeuten, in welchem Falle die Klassenmitten gemischte Zahlen mit dem Bruche $^1/_2$ tragen, oder es findet das Umgekehrte statt. Beträgt die Klassengröße 2 cm, so wird man die Wechselpunkte entweder mit den geraden oder den ungeraden Zahlen numerieren, die Klassenmitten tragen dann ungerade, beziehungsweise gerade Zahlen. Ist das Alter die Ordnungsgröße und hat man sich für ein Klassenintervall von 5 Jahren entschieden, so können die Wechselpunkte bei 0, 5, 10, 15, ... Jahren liegen, die Klassenmitten tragen dann die Bezeichnung 2,5; 7,5; 12,5; ...; oder man verlegt die Klassenmitten auf 0, 5, 10, ..., dann tragen die Wechselpunkte die Bezifferung —2,5; 2,5; 7,5; 12,5; ...; in diesem letzteren Falle gehört das erste Intervall zum Teil dem außermöglichen Gebiet an. In manchen Fällen gehören negative Werte mit zum Bereich des Kollektivs.

Auch eine einmal angenommene Lage der Klassenskala kann sich hinterdrein als abänderungsbedürftig erweisen.

c) Ist das Klassengerippe nach Weite und Lage festgestellt, so kommt es zur *Einreihung* der Glieder der Kollektivreihe in dasselbe und zur Bestimmung der Klassenhäufigkeiten. Dieser Vorgang richtet sich nach den geleisteten Vorarbeiten. Am einfachsten gestaltet er sich, wenn man zunächst die beobachteten Werte in eine steigende oder fallende Reihe ordnet (primäre Verteilungstafel). In dieser führen wiederholt vorkommende Argumentwerte die Wiederholungszahl neben sich. Die Klasseneinteilung vollzieht sich dann durch Einzeichnung von Grenzstrichen, und die Abzählung der Glieder zwischen je zwei Grenzstrichen liefert die Klassenhäufigkeiten.

Indessen ist bei einem umfangreichen Kollektiv die Umformung der Urliste in eine primäre Verteilungstafel eine zeitraubende und mühsame Arbeit. Man kann auch von der Urliste selbst Gebrauch machen, indem man ihre Glieder durch Striche bei den zutreffenden Klassen abbildet (Strichelungsverfahren). Um die nachfolgende Zählung zu erleichtern, empfiehlt es sich, die Striche zu je fünf zusammenzufassen, etwa so:

𝍸.

Zur Kontrolle dient die Summe der Klassenhäufigkeiten, sie muß mit dem im voraus festgestellten Umfang des Kollektivs übereinstimmen. Trifft das nicht zu, dann bleibt nur die Wiederholung des ganzen Verfahrens übrig.

Bei sehr umfangreichen Arbeiten dieser Art (Volks- und Berufszählungen, Wohnungszählungen u. a.) geht man so vor, daß für jedes Glied des Kollektivs eine Karte, *Zählkarte* genannt, angelegt wird, auf der die Argumentwerte für die verschiedenen Merkmale verzeichnet sind. Die Einreihung in Klassen geschieht durch Sortierung der Karten in einzelne Stöße (Päckchen), deren jedes einer Klasse entspricht; die Auszählung der Stöße gibt die Klassenhäufigkeiten. Die Kontrolle besteht hier in einem Durchprüfen der Stöße daraufhin, ob keine Karte verlegt ist. Bei sehr großen Kollektiven verwendet man zweckmäßig an Stelle von Zählkarten Lochkarten, auf denen die Argumentwerte der verschiedenen Merkmale durch Stanzen von Löchern zur Darstellung gebracht werden. Die Sortierung der Lochkarten in einzelne Stöße, von denen jeder ebenfalls einer Klasse entspricht, erfolgt mittels elektrisch betriebener Sortiermaschinen. Die Zählung der Lochkarten jedes Stoßes wird entweder durch Zählapparate an den Sortiermaschinen oder durch besondere Zählmaschinen bewirkt. Die Einreihung in ein im voraus festgesetztes Klassengerippe kann auch schon während der Aufnahme des Kollektivs geschehen; es fallen dann Urliste und primäre Verteilungstafel weg. Wird z. B. bei der Messung der Körpergröße von Rekruten die nächstliegende volle Zentimeterzahl angesagt, so ist damit die Mitte der betreffenden Zentimeterklasse bezeichnet und der Fall durch einen Strich an dieser Stelle abzubilden.

Eine Einzelfrage, die sich bei der Einreihung ergeben kann, muß noch besprochen werden. Wie sind Glieder zu behandeln, deren Argumentwert mit einem Wechselpunkt zusammenfällt, die also an der Grenze zweier Klassen liegen?

Ist der Argumentwert das Ergebnis einer Rechnung, z. B. eine Verhältniszahl, so kann mitunter durch Ausrechnung weiterer Dezimalstellen entschieden werden, in welche der beiden an den Wechselpunkt grenzenden Klassen das Glied

Tab. 13. Verteilungstafel mit der Klassengröße 1 dm.

X	z	X	z	X	z	X	z
55— 65	1	115—125	3	175—185	17,5	225—235	3
65— 75	1	125—135	5	185—195	13	235—245	2
75— 85	0	135—145	6	195—205	13,5	245—255	3
85— 95	0	145—155	11	205—215	6	255—265	1
95—105	1	155—165	10	215—225	7	265—275	1
105—115	3	165—175	17				125

Zwei Klassen sind leer; die Häufigkeitszahlen zeigen, namentlich im zweiten Teil, ein beständiges Auf- und Abschwanken, keine deutlich ausgeprägte Gesetzmäßigkeit.

In solchen Fällen versucht man es mit einer Zusammenziehung der Tafel auf einen kleineren Umfang, indem man — das ist der zweckmäßigste Vorgang — je zwei oder mehrere der bisherigen Klassen zu einer neuen Klasse vereinigt; das kann, wenn je v Klassen zusammengenommen werden, auf v verschiedene Arten geschehen, indem man entweder bei der 1., 2., ... oder v-ten Klasse beginnt, wobei man der Anfangsklasse so viele leere Klassen vorsetzt, als zu v fehlen; ähnlich am Schlusse. Man nennt eine so entstandene Verteilungstafel mit Bezug auf die ursprüngliche eine reduzierte Verteilungstafel, und spricht von verschiedenen Reduktionslagen je nachdem, wo man mit den neuen Klassen beginnt.

Wenn wir im vorliegenden Falle je zwei Klassen zusammenlegen, also zur Klassengröße 2 dm übergehen, so ergeben sich die folgenden zwei Tabellen:

Tab. 14. Reduzierte Verteilungstafeln mit der Klassengröße 2 dm.

Reduktionslage I.

X	z
55— 75	2
75— 95	0
95—115	4
115—135	8
135—155	17
155—175	27
175—195	30,5
195—215	19,5
215—235	10
235—255	5
255—275	2
	125

Reduktionslage II.

X	z
45— 65	1
65— 85	1
85—105	1
105—125	6
125—145	11
145—165	21
165—185	34,5
185—205	26,5
205—225	13
225—245	5
245—265	4
265—285	1
	125

Ein Auf- und Abschwanken findet nicht mehr statt; in beiden Tafeln steigen die z bis zu einem größten Wert (30,5, 34,5) an und fallen dann wieder ohne Rückschlag ab.

2) Gewichte von männlichen und weiblichen Neugeborenen[1]). In den Studienjahren 1902/03 und 1904/05 sind an der Gebärklinik in Bologna Wägungen an ausgetragenen Kindern beiderlei Geschlechts vorgenommen und in Urlisten veröffentlicht worden. Nachstehend sind die aus ihnen gewonnenen Verteilungstafeln, zuerst mit einer Klassengröße von 100 g, mitgeteilt.

[1]) C. Gini, Sulla variabilità dei due sessi, Cagliari 1910, p. 22—43.

Tab. 15. Verteilungstafel männlicher und weiblicher Neugeborener mit der Klassengröße 100 g.

X	x	z		X	x	z	
		Knaben	Mädchen			Knaben	Mädchen
1550—1650	1600	.	0,5	3150—3250	3200	30.5	27
1650—1750	1700	.	0,5	3250—3350	3300	28,5	17,5
1750—1850	1800	.	1	3350—3450	3400	23,5	16
1850—1950	1900	1	.	3450—3550	3500	18	19,5
1950—2050	2000	1,5	1	3550—3650	3600	18,5	13
2050—2150	2100	2,5	2	3650—3750	3700	13,5	11
2150—2250	2200	3	3	3750—3850	3800	13	8
2250—2350	2300	4,5	9	3850—3950	3900	6	3
2350—2450	2400	1,5	6,5	3950—4050	4000	4	2,5
2450—2550	2500	3	8	4050—4150	4100	5	1
2550—2650	2600	8	8,5	4150—4250	4200	6,5	1
2650—2750	2700	14	18	4250—4350	4300	3,5	.
2750—2850	2800	14,5	17	4350—4450	4400	2,5	.
2850—2950	2900	14	18	4450—4550	4500	0,5	.
2950—3050	3000	17	25	4550—4650	4600	3	1
3050—3150	3100	26	30,5	4650—4750	4700	1	.
						288	269

Die beiden Verteilungen zeigen trotz der Unregelmäßigkeit im einzelnen doch schon charakteristische Unterschiede, deren rechnerische Erfassung erst später erfolgen kann.

Auch bei einer Erhöhung der Klassengröße auf 200 g verschwinden die Schwankungen nicht vollständig; aber bei 300 g tritt dies schon ein, wie aus den folgenden reduzierten Tafeln (mit ungleicher Reduktionslage) hervorgeht.

Tab. 16. Reduzierte Verteilungstafeln mit der Klassengröße 300 g.

Knaben.

X	x	z
1850—2150	2000	5
2150—2450	2300	9
2450—2750	2600	25
2750—3050	2900	45,5
3050—3350	3200	85
3350—3650	3500	60
3650—3950	3800	32,5
3950—4250	4100	15,5
4250—4550	4400	6,5
4550—4850	4700	4
		288

Mädchen.

X	x	z
1450—1750	1600	1
1750—2050	1900	2
2050—2350	2200	14
2350—2650	2500	23
2650—2950	2800	53
2950—3250	3100	82,5
3250—3550	3400	53
3550—3850	3700	32
3850—4150	4000	6,5
4150—4450	4300	1
4450—4750	4600	1
		269

3) Schädelindex umbrischer Rekruten[1]).

In den Assentjahren 1859 bis 1863 sind an 6209 aus Umbrien stammenden Rekruten Schädelmessungen vorgenommen worden: für den daraus abgeleiteten Schädelindex, durch die nächstliegende ganze Zahl ausgedrückt, ergab sich folgende Verteilung.

Tab. 17. Verteilungstafel der Schädelindizes umbrischer Rekruten.

x	z	x	z	x	z
68	1	79	370	90	149
69	1	80	299	91	155
70	2	81	391	92	111
71	4	82	550	93	54
72	8	83	573	94	57
73	13	84	722	95	23
74	28	85	456	96	9
75	45	86	608	97	10
76	95	87	369	98	3
77	152	88	316		
78	188	89	447		6209

Die Klassengrenzen sind 67,5—68,5, 68,5—69,5 usw.

Die Klassenhäufigkeiten zeigen namentlich im Kern der Tafel beträchtliche Schwankungen, und man wird versuchen, sie durch Reduktion zu beseitigen. Entscheidet man sich, je drei Klassen zusammenzulegen, so kann mit 67 oder 68 oder 69 begonnen werden; man hat dann 2, 1, 0 leere Klassen vorzulegen und am Ende 0, 1, 2 leere Klassen anzufügen. Unterscheidet man die aufgezählten drei Reduktionslagen durch I, II, III, so erhält man folgendes Bild.

Tab. 18. Reduzierte Verteilungstafeln in drei verschiedenen Lagen.

I		II		III	
x	z	x	z	x	z
67	1	68	2	69	4
70	7	71	14	72	25
73	49	74	86	75	168
76	292	77	435	78	710
79	857	80	1060	81	1240
82	1514	83	1845	84	1751
85	1786	86	1433	87	1293
88	1132	89	912	90	751
91	415	92	320	93	222
94	134	95	89	96	42
97	22	98	13	99	3
	6209		6209		6209

[1]) C. Gini, Variabilità e mutabilità, Bologna 1912, p. 30.

Die Schwankungen sind durchwegs verschwunden. Aber die Tafel III ist den andern an Regelmäßigkeit überlegen und zeigt im Kern, von $x = 78$ bis $x = 90$, den höchsten Grad von Symmetrie, die bei den andern weniger deutlich hervortritt.

Wenn einer Klasse die Klassenmitte als Argument zugewiesen wird, so liegt darin eine Ungenauigkeit; richtiger wäre es, das arithmetische Mittel der in die Klasse fallenden Argumentwerte zu nehmen; das könnte bei der ursprünglichen Verteilungstafel nur auf Grund der Urliste geschehen, bei einer reduzierten Tafel hätte es mit Benützung der z als Gewichte zu erfolgen. Wir wollen an dem vorliegenden Beispiel zeigen, daß eine solche Verschärfung der Rechnung keine wesentlichen Verschiebungen nach sich zieht.

In I ergibt sich auf diesem Wege für die Klasse 85 das Argument

$$\frac{722 \cdot 84 + 456 \cdot 85 + 608 \cdot 86}{1786} = 84{,}94,$$

in II für die Klasse 80

$$\frac{370 \cdot 79 + 299 \cdot 80 + 391 \cdot 81}{1060} = 80{,}02,$$

n III für die Klasse 81

$$\frac{299 \cdot 80 + 391 \cdot 81 + 550 \cdot 82}{1240} = 81{,}20.$$

Am größten wird die Ungenauigkeit an den Enden, besonders wenn leere Klassen zugefügt werden mußten. So wären in unseren drei Fällen die schärfer gerechneten Argumentwerte der Endklassen 96,7; 97,2 und 98.

26. Die Größengleichheit der Klassen ist ein wesentliches Erfordernis, dem in statistischen Veröffentlichungen nicht immer durchaus entsprochen ist; das erschwert die Beurteilung und verhindert die Anwendung der später zu entwickelnden Rechnungsweisen. Besonders häufig kommt es vor, daß am Schlusse einer Tafel eine Zusammenfassung des Restes ohne Angabe der oberen Grenze erfolgt; so schließt z. B. der Altersaufbau einer Bevölkerung gewöhnlich mit der unbegrenzten Klasse „hundert Jahre und darüber". Im Innern der Tafel macht sich der Wechsel in der Klassengröße meist durch ein plötzliches Anwachsen oder Fallen in der Häufigkeitszahl bemerkbar. Ohne Kenntnis des Urmaterials ist die Umformung einer solchen Tafel in eine gleichklassige undurchführbar; Zerlegung großer Klassen in kleinere unter gleichmäßiger Aufteilung der Häufigkeit ist nur ein ungenauer Notbehelf.

Ein Beispiel für eine solche ungleichmäßige Klasseneinteilung gibt die umstehende finanzstatistische Verteilungstafel 19.

Nicht weniger als fünfmal wechselt die Klassengröße, und beim zweiten und vierten Wechsel verrät sich dies schon an der Zahl z. Die Beibehaltung der anfänglichen Klassengröße würde zu einer sehr langen Tafel führen, in deren oberem Teil Klassen unbesetzt blieben. Die in den letzten beiden Spalten bis 4200 vorgenommene gleichmäßige Aufteilung entfernt sich sicher von der Wirklichkeit und hätte, je höher die Stufen werden, um so weniger Sinn.

Tab. 19. Die steuerbelasteten Lohnsteuerpflichtigen im Deutschen Reich 1928.[1])

Einkommensgruppen X	Steuerbelastete z	Gleiche Einkommensgruppen X	Gleichmäßige Aufteilung z
bis 1200 RM.	4 763 297	bis 1200 RM.	4 763 297
über 1200 „ 1500 „	1 476 822	über 1200 „ 1500 „	1 476 822
„ 1500 „ 1800 „	1 422 182	„ 1500 „ 1800 „	1 422 182
„ 1800 „ 2100 „	1 315 660	„ 1800 „ 2100 „	1 315 660
„ 2100 „ 2400 „	1 072 713	„ 2100 „ 2400 „	1 072 713
„ 2400 „ 3000 „	1 389 866	„ 2400 „ 2700 „	694 933
„ 3000 „ 3600 „	705 020	„ 2700 „ 3000 „	694 933
„ 3600 „ 4200 „	415 537	„ 3000 „ 3300 „	352 510
„ 4200 „ 5000 „	353 351	„ 3300 „ 3600 „	352 510
„ 5000 „ 6500 „	382 451	„ 3600 „ 3900 „	207 768,5
„ 6500 „ 8000 „	154 802	„ 3900 „ 4200 „	207 768,5
„ 8000 „	38 764		.
	13 490 465		.

Die Ungleichmäßigkeit der Klassen ist indessen mitunter geboten, um kennzeichnende Züge der Verteilung zum Ausdruck zu bringen, die sonst bei gleichmäßiger, aber zu grober Gliederung verloren gingen. So erhielte man z. B. nur ein unvollkommenes Bild von dem Verlauf der Sterblichkeit, wenn man die Gesamtheit der im ersten Lebensjahre gestorbenen Kinder nicht feiner nach dem Alter gliedern würde.

Tab. 20. Die Gestorbenen des ersten Lebensjahres in Sachsen 1933.[2])

Alter	Zahl der Gestorbenen	Alter	Zahl der Gestorbenen
unter 1 Tag	911	2—3 Wochen	157
1—2 Tage	366	3 Wochen bis 1 Monat..	113
2—3 „	170	1—2 Monate	294
3—4 „	112	2—3 „	227
4—5 „	46	3—6 „	510
5—6 „	47	6—9 „	314
6—7 „	38	9—12 „	201
1—2 Wochen	203		

[1]) Statistik des Deutschen Reichs, Band 378, S. 23.

[2]) Statistisches Jahrbuch für das Land Sachsen, 50. Ausgabe 1931/34, S. 74.

Erst aus dieser feineren Aufteilung der im ersten Lebensjahr gestorbenen Kinder nach dem Alter erkennt man die hohe Sterblichkeit am Anfang des Lebens, vor allem in den ersten beiden Lebenstagen.

Ein sehr bemerkenswertes Beispiel dieser Art bietet auch die Verteilung der Sterbefälle an Diphtherie auf die verschiedenen Alter; die folgende Tabelle gibt sie nach den Beobachtungen in England und Wales während des Jahrzehnts 1891 bis 1900.

Bei einer gleichmäßigen Klasseneinteilung wäre die charakteristische Erscheinung des starken Ansteigens der Diphtheriesterblichkeit zu ihrem Maximum im vierten Lebensjahr der Wahrnehmung entgangen, es hätte sich nur die gewaltige Gesamtsterblichkeit des ersten Jahrfünfts mit 49 479 Fällen, das ist mit 61,3 % der sämtlichen Opfer dieser Krankheit, herausgestellt. Die dritte Spalte, welche die jährlichen Quoten, vom 6. Jahre an die durchschnittlichen, angibt, läßt deutlicher als die zweite Spalte die rasche Abnahme der Sterblichkeit nach dem ersten Jahrfünft erkennen.

Tab. 21. Die Diphtheriesterbefälle in England und Wales 1891 bis 1900.[1])

Alter in Jahren X	Zahl der Todesfälle z	Einjährige Quote
0— 1	4 186	4 186
1— 2	10 491	10 491
2— 3	11 218	11 218
3— 4	12 390	12 390
4— 5	11 194	11 194
5—10	23 348	4 670
10—15	4 092	818
15—20	1 123	225
20—25	585	117
25—35	786	79
35—45	512	51
45—55	324	32
55—65	260	26
65—75	127	13
75 und darüber	35	unbestimmt
	80 671	

Auch gegenwärtig zeigt die Diphtheriesterblichkeit in den ersten Lebensjahren eine Zunahme. Nach der von E. Meier[2]) auf Grund der preußischen Zahlen durchgeführten Berechnung entfielen in den Jahren 1927 bis 1931 auf 100 000 gleichaltrig Lebende an Diphtherie Gestorbene:

Tab. 22.

Alter in Jahren	Sterbeziffer	Alter in Jahren	Sterbeziffer
0— 1	28,0	10—15	6,47
1— 2	40,3	15—30	0,53
2— 3	41,3	30—60	0,45
3— 5	50,3	60—70	0,37
5—10	40,2	70 und darüber	0,22

[1]) Vgl. G. U. Yule, An Introduction to the Theory of Statistics, London 1932, p. 98.

[2]) E. Meier, Die Altersverteilung der neuen Diphtheriewelle, Reichsgesundheitsblatt, 10. Jahrgang, 1935, S. 24. Für die Altersgruppen von 15 Jahren ab sind die Sterbeziffern nach dem Statistischen Jahrbuch für den Freistaat Preußen, 25. Band 1929, S. 79; 26. Band 1930, S. 69; 27. Band 1931, S. 42 und 117; 28. Band 1932, S. 63 und 29. Band 1933, S. 49 berechnet worden. Vgl. K. Pohlen, Gesundheitsstatist. Auskunftsbuch 1936, S. 162.

27. Eine längere Zahlenreihe ist in ihrer Gesamtheit schwer zu erfassen, namentlich, wenn sie aus großen Zahlen besteht. Viel anschaulicher ist ihre geometrische Darstellung: sie läßt den allgemeinen Verlauf und etwaige Besonderheiten mit einem Blick überschauen.

Für Verteilungstafeln sind zwei Darstellungsweisen üblich.

Die eine besteht darin, daß man auf einer Achse das Klassengerippe nach einem gewählten Maßstab aufträgt und in den Mittelpunkten der Klassenintervalle Lote zu der Achse errichtet, deren Länge die Häufigkeit der betreffenden Klasse wiedergibt; der zugehörige Maßstab erfordert einige Überlegung. Um dem Auge einen besseren Halt zu geben, verbindet man die Endpunkte der Lote durch einen gebrochenen Linienzug und erhält so ein die Verteilungstafel veranschaulichendes Häufigkeitspolygon. Seine Endpunkte liegen in der Achse und fallen in die Mitten der der ersten vorausgehenden und der letzten nachfolgenden leeren Klasse.

Fig. 1. Häufigkeitspolygon der Schädelindizes umbrischer Rekruten.

Fig. 1 zeigt das Häufigkeitspolygon, gehörig zur Verteilungstafel 17 der Schädelindizes umbrischer Rekruten (Art. 25, 3).

Bei der zweiten Darstellungsweise werden über den Klassenintervallen Rechtecke errichtet, deren Höhen ebenso bemessen sind wie die früheren Lote, die nunmehr als Symmetrielinien der Rechtecke anzusprechen sind.

Auf diese Art entsteht ein nach obenhin treppenartig begrenztes Bild, das wir als Staffelbild der Verteilung bezeichnen wollen. Sein Zusammenhang mit dem Häufigkeitspolygon ist ein einfacher: letzteres ergibt sich aus dem Staffelbild, indem man die Endpunkte der erwähnten Symmetrieachsen der Reihe nach verbindet.

Fig. 2 ist das Staffelbild der Verteilung der Höhen neunjähriger Kiefern in der Reduktionslage II (Tab. 14, Art. 25, 1); ihm ist auch das Häufigkeitspolygon eingezeichnet.

So verschieden der Anblick eines Häufigkeitspolygons und eines Staffelbildes ist, in einem stimmen beide miteinander überein: in der Fläche, die sie mit der Grundlinie begrenzen. Dies ist aus der Fig. 2 unmittelbar ersichtlich; die Dreiecke, die das Polygon vom Staffelbild abschneidet und zu ihm hinzufügt, gleichen sich paarweise aus. Somit stellt die Fläche, welche durch die Grundlinie und das Polygon begrenzt wird, den Um-

Fig. 2. Staffelbild und Häufigkeitspolygon der Höhen neunjähriger Kiefern.

fang des Kollektivs dar. Es ist daher notwendig, den Wert eines Kollektivgliedes in Flächeneinheiten anzugeben, und das hängt von den angewendeten Maßstäben ab. Angenommen, daß im vorliegenden Falle die Klasse durch 8 mm, ein Bäumchen auf dem Höhenmaßstab durch 2 mm dargestellt ist, so entspricht einem Kollektivglied eine Fläche von $8 \times 2 = 16\,\text{mm}^2$, und es beträgt die Fläche von Staffelbild und Polygon $125 \times 16 = 2000\,\text{mm}^2$.

Während jedoch das über einem Klassenintervall ruhende Rechteck in seiner Fläche die der Klasse entsprechende Häufigkeit darstellt, gilt dies von dem zugehörigen Streifen des Häufigkeitspolygons im allgemeinen nicht, sondern nur dann, wenn seine obere Begrenzung nicht eine gebrochene, sondern eine durchlaufende Gerade ist, was wieder nur dann stattfindet, wenn der in der Mitte des Streifens befindliche Eckpunkt mit den beiderseits benachbarten in einer Linie liegt. In anderen Fällen ist der Streifen größer oder kleiner, je nachdem das Polygon an der Stelle einen konkaven oder konvexen Winkel (von der Abszissenachse aus gesehen) hat.

Das Staffelbild trägt der Anschauung Rechnung, die Glieder einer Klasse seien über das Klassenintervall gleichmäßig verteilt, so daß auf einen Teil des Intervalls der entsprechende Teil der Häufigkeit entfällt. Das Häufigkeitspolygon entspricht in diesem Punkte den Tatsachen besser, indem es die Häufigkeit von der Mitte aus nach der einen Seite zu-, nach der andern abnehmend und nur in einem Höchstpunkt nach beiden Seiten abnehmend darstellt. Doch wird bei den später vorzunehmenden Rechnungen an der gleichmäßigen Verteilung festgehalten.

28. Die bisherige Darstellung der Verteilung eines Kollektivs hat den Nachteil, daß mehrere Verteilungen untereinander nicht vergleichbar sind, wenn sie nicht zu Kollektiven gleicher Umfänge gehören. Das aber läßt sich im allgemeinen nicht nach Belieben herbeiführen. Man kann aber dasselbe Ziel auch bei ungleichen Umfängen erreichen, indem man alle Kollektive auf einen festen Umfang umformt. Als solcher wird eine der Zahlen 1, 100, 1000 oder eine noch höhere Potenz von 10 verwendet.

Neben den Begriff der absoluten Häufigkeit tritt so der neue der relativen Häufigkeit.

Um auf die Zahl 1 zu reduzieren, hat man die absoluten Häufigkeiten durch den Umfang N zu dividieren, und um dann auf 100, 1000 überzugehen, bedarf es noch der Multiplikation mit diesen Zahlen.

Mit Bezug auf die zeichnerische Veranschaulichung bedeutet dies so viel, daß man die ganze Fläche des Häufigkeitsdiagramms als Maß verwendet und ihr entweder die Zahl 1 oder 100 oder 1000 zuordnet. An dieser Fläche gemessen gibt dann jeder Klassenstreifen die relative Häufigkeit der betreffenden Klasse.

Neben der Klassenhäufigkeit findet wegen ihrer Anschaulichkeit die Häufigkeit bis zu einem bestimmten Wechselpunkt vielfache Anwendung; sie bezeichnet jenen Teil des Kollektivs, bei dem das Argument eine bestimmte Grenze nicht überschreitet, und ergibt sich durch Summierung der Häufigkeiten der bis zu dieser Grenze reichenden Klassen. So stellt sich jeder Verteilungstafel ihre Summentafel an die Seite, und es muß daran gedacht werden, daß ihre Glieder Geltungsbereiche haben, die mit einem Wechselpunkt und nicht mit einer Klassenmitte abschließen.

Nachstehend ist neben die Verteilungstafel 13 (Art. 25, 1) der Höhen neunjähriger Kiefern mit der Klassengröße 1 dm die zugehörige Summentafel gestellt; im Anschluß daran ist an der Summentafel die Reduktion auf 1 durchgeführt und die Tafel der auf dieselbe Basis zurückgeführten relativen Klassenhäufigkeiten als Differenzentafel der vorausgehenden abgeleitet. Dieser Vorgang hat vor einer unmittelbaren Berechnung der relativen Klassenhäufigkeiten den Vorzug, daß deren Summe, wie es sein soll, den genauen Wert 1 erhält, was bei dem direkten Verfahren wegen der notwendigen Abrundungen im allgemeinen nicht zu erreichen wäre.

Tab. 23. Summentafel etc. der Höhen neunjähriger Kiefern.

X	z	Summentafel	Relative Summentafel	Relative Klassenhäufigkeit
55— 65	1	1	0,008	0,008
65— 75	1	2	0,016	0,008
75— 85	0	2	0,016	0,000
85— 95	0	2	0,016	0,000
95—105	1	3	0,024	0,008
105—115	3	6	0,048	0,024
115—125	3	9	0,072	0,024
125—135	5	14	0,112	0,040
135—145	6	20	0,160	0,048
145—155	11	31	0,248	0,088
155—165	10	41	0,328	0,080
165—175	17	58	0,464	0,136
175—185	17,5	75,5	0,604	0,140
185—195	13	88,5	0,708	0,104
195—205	13,5	102	0,816	0,108
205—215	6	108	0,864	0,048
215—225	7	115	0,920	0,056
225—235	3	118	0,944	0,024
235—245	2	120	0,960	0,016
245-255	3	123	0,984	0,024
255—265	1	124	0,992	0,008
265—275	1	125	1,000	0,008
	125			1,000

Diese Tafel besagt beispielsweise in ihrer zehnten Zeile, daß 11 Pflanzen eine Höhe von 15 dm aufwiesen, d. h. eine Höhe, die von diesem Maße nicht über 5 cm nach oben oder unten abwich; daß 31 Pflanzen die Höhe von 155 cm nicht überschritten; daß diese 0,248 oder 24,8 % aller Bäumchen ausmachten; daß der Klasse 145—155 cm 0,088 oder 8,8 % aller Bäumchen angehörten.

Der Verteilungstafel gegenüber hat die Summentafel den Vorzug, daß ihre Glieder ständig wachsen, daß Schwankungen, die in der ersten etwa vorkommen, in der Summentafel verschwinden.

Um aus dem Häufigkeitspolygon oder dem Staffelbild das Summenpolygon zu erhalten, hat man in den Klassenenden Lote zu errichten, die gleich sind der Summe der vorhergehenden Eckpunktsordinaten oder der vorhergehenden Rechteckshöhen, und ihre Endpunkte durch gerade Linien zu verbinden. Der Sinn des Summenpolygons ist der folgende: Irgend eine Ordinate desselben, als Höhe eines Rechtecks mit der Klassengröße als Basis genommen, liefert in der Fläche dieses Rechtecks, nach den zugrunde gelegten Maßstäben gemessen, die absolute Häufigkeit der Glieder, deren Argument über den Fußpunkt der gewählten Ordinate nicht hinausgeht.

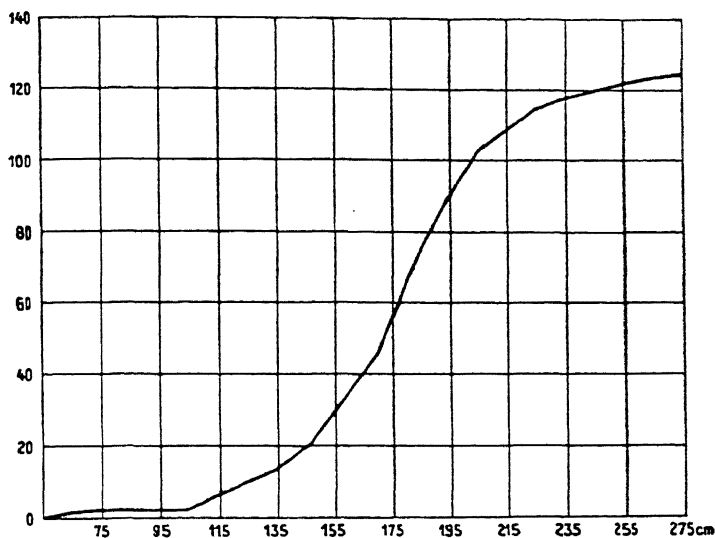



Fig. 3. Summenpolygon der Höhen neunjähriger Kiefern.

Fig. 3 gibt ein Bild der vorstehenden Summentafel. Wegen der gemeinsamen Basis der Rechtecke kann die Häufigkeit auch an dem Höhenmaßstab selbst abgelesen werden und wird zur relativen Häufigkeit, wenn man sie an der Endordinate als Einheit mißt.

29. Mit beständig wachsendem Umfang eines Kollektivs wird man einerseits die Klassengröße herabsetzen können und anderseits wird sich die Besetzung der Klassen verstärken. Durch beide Umstände nähert sich das Häufigkeitspolygon immer mehr einer glatt verlaufenden krummen Linie, weist das Staffelbild immer mehr auf eine solche hin. Man gelangt so zu dem Begriff der Häufigkeitskurve, von der man annimmt, daß sie eine mehr oder weniger vollkommene Verwirklichung einer idealen Kurve sei, die der Materie, von der das Kollektiv eine Probe darstellt, eigentümlich ist. Man kann nämlich annehmen, und Erfahrungen sprechen dafür, daß einer jeden Materie eine Häufigkeitskurve entspricht, die für sie ebenso kennzeichnend ist wie etwa der Habitus für eine Pflanze, das spezifische Gewicht für ein Mineral.

Bei der unübersehbaren Mannigfaltigkeit von Materien, die einer kollektiven Behandlung zugänglich sind, sollte man von vornherein erwarten, daß die Zahl der Formen von Häufigkeitslinien eine sehr große sein werde. Indessen trifft diese Erwartung nicht zu, vielmehr lassen sich die vorkommenden Formen der Hauptsache nach einer kleinen Zahl von Typen unterordnen.

Das gilt allerdings nur so lange, als man es mit gleichförmigen, mit homogenen Materien zu tun hat. Anders verhält es sich mit Kollektiven gemischter Zusammensetzung. Hier können durch Zusammentreffen verschiedener Verteilungen die mannigfachsten Häufigkeitskurven entstehen. Außerdem wirken auch störende Einflüsse auf die Gestaltung der Kurve. Es hat daher nicht alles, was an den Häufigkeitskurven in die Erscheinung tritt, tiefere Bedeutung, und nur wiederholte Untersuchungen an einer und derselben Materie können das Wesentliche vom Unwesentlichen trennen. Es machte sich oft eine Überschätzung der Einzelheiten des Kurvenverlaufs und damit zusammenhängend eine zu weit gehende Klassifikation der Kurventypen bemerkbar, mit der aber nur ein scheinbarer Gewinn an Erkenntnis erreicht wurde.

Wenn sich die Theorie mit gewissen idealen Häufigkeitskurven eingehend beschäftigt, so verfolgt sie den gleichen Weg, den viele Zweige der reinen Wissenschaft einschlagen; so stellt die Geometrie ihre Untersuchungen an ideellen Raumgebilden an mit dem Bewußtsein, daß sie nirgends eine vollkommene Verwirklichung finden; und doch weiß man, welch eine wertvolle Grundlage für die Erforschung der Wirklichkeit sie damit schafft.

30. Den Ausgangspunkt für die Untersuchung, Beurteilung und Klassifikation von Verteilungen bildet eine ideale Verteilung nach einer Häufigkeitskurve von der Gleichungsform

$$y = \frac{h}{\sqrt{\pi}} e^{-h^2 (X-A)^2} \tag{1}$$

Wir wollen sie die normale Häufigkeitskurve[1]) nennen. Was an ihr in erster Linie hervortritt, ist die Symmetrie der Verteilung in Bezug auf einen bestimmten Ausgangswert A, in welchem y sein Maximum, die Häufigkeit also den größten Wert erreicht. Von da ab nimmt sie nach beiden Seiten gleichförmig ab. Schon daran, daß sich die Kurve ins Unendliche erstreckt, während die Verteilung eines jeden, noch so umfangreichen Kollektivs nur eine endliche Erstreckung aufweisen kann, ist zu erkennen, daß die Kurve nur ideellen Charakten haben kann; daß jedoch die erwähnte Divergenz praktisch belanglos ist, liegt in dem eigenartigen Charakter der Funktion (1), in ihrer außerordentlich raschen Abnahme bei wachsendem Betrage des Arguments X. Ihrer Gestalt nach kann die Normalkurve, wie sie kurz genannt werden soll, mit dem Profil einer Glocke verglichen werden, wie Fig. 4 zeigt.

Verlegt man den Ausgangspunkt der Argumentzählung nach A, mit andern Worten, rechnet man statt mit dem ursprünglichen Argumentwert X mit seiner Abweichung x von A, so tritt an die Stelle von (1) die einfachere Gleichung

$$y = \frac{h}{\sqrt{\pi}} e^{-h^2 x^2} \tag{2}$$

[1]) Die Gleichungsform (1) wird in Art. 117 (S. 278 u. f.) entwickelt.

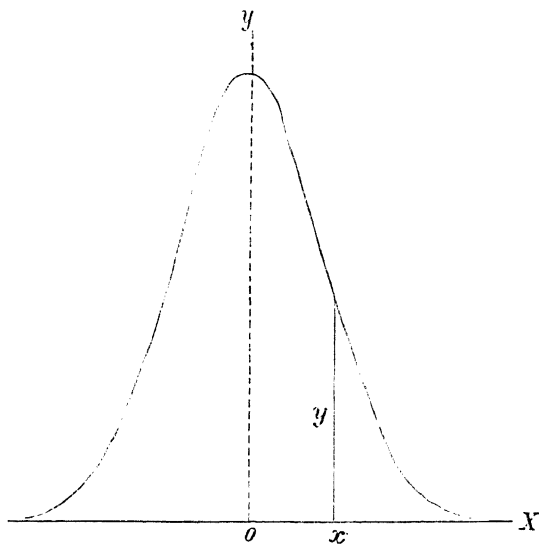



Fig. 4. Normale Häufigkeitskurve.

Der wesentliche Unterschied der Gleichungen (1) und (2) liegt darin, daß die erste zwei unbestimmte Parameter, A und h, die zweite nur mehr einen Parameter, h, hat. Der Parameter A hat nur auf die Lage der Kurve zum Koordinatensystem Einfluß, der Parameter h bestimmt ihre Gestalt, wie schon daraus zu erkennen ist, daß die maximale Ordinate (für $x = 0$) $y_0 = \frac{h}{\sqrt{\pi}}$ von h abhängt.

Die Gleichungen (1), (2) sind so gestaltet, daß sich für die zwischen der Kurve und der Abszissenachse gelegene Fläche, entsprechend der Auffassung aller relativen Häufigkeitsdarstellungen, der Wert 1 ergibt, denn es ist

$$\frac{h}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-h^2(X-A)^2}\, dX = \frac{h}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-h^2x^2}\, dx = \frac{2}{\sqrt{\pi}} \int_0^{\infty} e^{-t^2}\, dt = 1.$$

Verteilungen, welche sich der Normalkurve mit hinreichender Genauigkeit anpassen lassen, gehören zu den Ausnahmen und sind am ehesten bei anthropologischen, überhaupt der belebten Natur entnommenen Kollektiven zu treffen. Schon die Symmetrie, die dabei vorausgesetzt wird, ist eine Eigenschaft, die niemals rein zum Ausdruck kommt, schon infolge der stets mitspielenden zufälligen Störungen: aber die Abweichungen davon können immerhin so geringfügig und so verteilt sein, daß man sich zu dem Ausspruche berechtigt fühlen kann, hinter dem endlichen Kollektiv stecke ein Gegenstand, in dessen Natur symmetrische Verteilung begründet ist.

Die Bedeutung der Normalkurve geht aber über diese Fälle hinaus. Die Größenbeziehungen, die sich aus ihr ergeben, bilden wertvolle Anhaltspunkte für die Beurteilung von Verteilungen, die von „normalen“ abweichen.

Als Beispiel einer Verteilung, von der man mit guter Berechtigung behaupten kann, daß ihr Symmetrie zugrunde liege, führen wir die von A. Quetelet[1]) erhobenen Brustumfänge von 1516 belgischen Soldaten an und stellen sie durch die Verteilungstafel und das Häufigkeitspolygon dar.

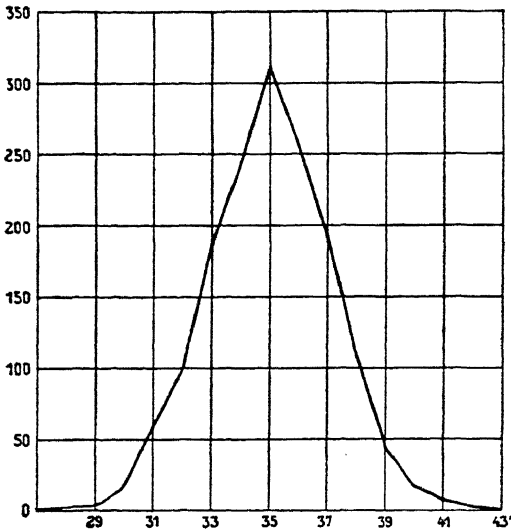



Fig. 5. Häufigkeitspolygon der Brustumfänge belgischer Soldaten.

Tab. 24.

x in belg. Zoll (″)	z
28	2
29	4
30	17
31	55
32	102
33	180
34	242
35	310
36	251
37	181
38	103
39	42
40	19
41	6
42	2
	1516

Die Messung von Körperhöhen und Brustumfängen wurde ursprünglich bei Männern bestimmter Altersklassen aus militärischen Gründen vorgenommen, später, seit dem Auftreten A. Quetelets, gesellten sich anthropologische und andere Interessen dazu, und es wurden auch andere, den menschlichen Körper betreffende Größen, so das Gewicht, verschiedene Kopfdimensionen u. a., nunmehr bei beiden Geschlechtern und in allen Altern in die Untersuchung einbezogen[2]). Schließlich

[1]) A. Quetelet, Recherches sur la loi de la croissance de l'homme. Nouveaux Mémoires de l'Académie de Bruxelles, t. VII, 1831.

[2]) Vgl. hierzu A. Quetelet, Anthropométrie ou mesure des différentes facultés de l'homme, Brüssel 1870. G. Fechner (Kollektivmaßlehre. Leipzig 1897, S. 102) hat Messungen an 450 europäischen Männerschädeln statistisch bearbeitet. H. Rautmann (Untersuchungen über die Norm, ihre Bedeutung und Bestimmung. Jena 1921, Tabellen-Beilagen) untersuchte mittels Messungen die Beziehungen der Körpergröße zu Körpergewicht, Brustumfang, Brustspielraum, Herzgröße, Pulszahl und Blutdruck. Den Zusammenhang zwischen Kopflänge und Kopfbreite unterzog J. Linders (Zur Kenntnis der Kopfmaße in Schweden. Medderlande Frau Lunds Astronomiska Observatorium, Ser. II Nr. 50a, 1927) an 2037 jungen Schweden einer Untersuchung. Die für anthropometrische Forschungen in Betracht kommenden mathematisch-statistischen Methoden sind in neuerer Zeit zusammenfassend von E. Weber (Einführung in die Variations- und Erblichkeits-Statistik. München 1935) dargestellt worden. Ferner hat F. Ringleb (Mathemathische Methoden der Biologie, insbesondere der Vererbungs-

kam noch ein praktisches Interesse hinzu, seit man erkannt hat, daß Körperhöhe, Brustumfang und Körpergewicht Faktoren sind, die auf die Sterblichkeit einen erkennbaren Einfluß üben. Damit ist die Anthropometrie zu einem wichtigen Gegenstand der Versicherungsmedizin und der Versicherungstechnik geworden. Dort, wo das Versicherungswesen in riesenhaftem Maßstabe betrieben wird, wie in Amerika, hat dies zur Aufstellung von different behandelten Risikenklassen geführt; aber auch bei engeren Verhältnissen können die anthropometrischen Faktoren nicht mehr übergangen, müssen vielmehr bei der Bewertung der Risiken mit zu Rate gezogen werden[1]).

31. Die Symmetrie in der Verteilung eines Kollektivs ist als ein Ausnahmefall anzusehen; die Regel bildet Asymmetrie, die der verschiedensten Grade fähig ist. Neben die symmetrische normale Häufigkeitskurve stellt sich also eine unbegrenzte Zahl asymmetrischer Kurven, welche mit der ersten das eine gemein haben, daß sie von einem Gipfelpunkte nach beiden Seiten abfallen, nur geschieht das auf der einen Seite rascher als auf der andern. Je nachdem der steiler abfallende Zweig der Kurve nach der Seite der großen Argumentwerte oder nach der Seite der kleinen liegt, hat man zwei Arten der Asymmetrie zu unterscheiden, die erste als die rechtsseitige, die zweite als die linksseitige.

Bei einem unstetigen Kollektiv kann eine beobachtete Asymmetrie entweder der Ausdruck einer wirklich vorhandenen (echte Asymmetrie) oder, wenn sie schwachen Grades ist, eine Folge zufälliger Störungen sein, was insbesondere bei unzureichendem Umfang eintreten kann. Bei einem stetigen Kollektiv ist noch ein drittes möglich; hier kann eine mäßige Asymmetrie auch aus der Art der Klasseneinteilung entspringen und verschwinden, wenn man zu einer anderen Einteilung übergeht. Darum sollte man in jedem Falle versuchen, ob sich eine sich zeigende schwache Asymmetrie nicht auf diesem Wege beseitigen läßt. Ob es sich um eine echte, also für die Materie kennzeichnende oder nur um scheinbare Asymmetrie handelt, kann bei schwachen Graden nur durch wiederholte Prüfung von genügend umfangreichen Kollektiven der betreffenden Art erforscht werden.

Fechner[2]) hat die Idealkurve einer asymmetrischen Verteilung aus zwei Ästen zusammengesetzt, die zwei verschieden gestalteten Normalkurven entnommen waren

lehre und der Rassenforschung. Leipzig 1937) weiterführende Berechnungen unter Anwendung mathematischer Methoden an verschiedenen anthropometrischen Kollektiven angestellt.

[1]) In Deutschland hat Karup (IV. Internationaler Kongreß für Versicherungsmedizin. Berlin 1906, S. 46 u. f.) als erster anthropometrische Messungen an den männlichen Zugängen der Gothaer Lebensversicherungsbank a. G. aus den Jahren 1881—1904 statistisch bearbeitet. Mit amerikanischen Arbeiten auf diesem Gebiete beschäftigt sich G. Bohlmann in dem Aufsatz: Anthropometrie und Lebensversicherung (Zeitschrift für die gesamte Versicherungs-Wissenschaft. Band 14, 1914, S. 743 u. f.). H. Wulkow (Die Körpermaße der Lebensversicherten. Berlin 1936) untersuchte an Hand des Materials der von 50 Gesellschaften in den Jahren 1930—1932 mit ärztlicher Untersuchung abgeschlossenen Großlebensversicherungen die Abhängigkeit der Körpermaße vom Alter und von der Körperhöhe. Einen Überblick über die neueren anthropometrischen Messungen für die Zwecke der Lebensversicherung gibt G. Florschütz in dem Aufsatz: Die Körpermaße der Lebensversicherten (Zeitschrift für die gesamte Versicherungs-Wissenschaft, 36. Bd. 1936, S. 339 u. f.)

[2]) G. Th. Fechner, Kollektivmaßlehre, Leipzig 1897, S. 294 u. f.

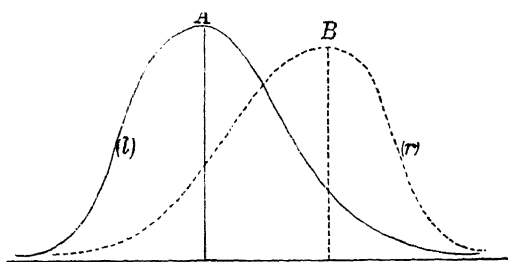



Fig. 6.
Rechtsseitig (r) und linksseitig (l) asymmetrische

und im gemeinsamen Scheitel zusammengefügt wurden. Pearson[1]) hat eine asymmetrische Häufigkeitskurve, die wie die Normalkurve ins Unendliche sich erstreckt, durch eine einheitliche Gleichung dargestellt. In Fig. 6 ist die allgemeine Form einer rechts- und linksseitig asymmetrischen Häufigkeitskurve angedeutet.

Die folgenden drei Beispiele deutlich asymmetrischer Verteilungen werden noch zu einigen Bemerkungen Anlaß geben.

1) K. Marbe[2]) hat zum Zwecke verschiedener biologischer Untersuchungen Erbsen bestimmter Arten angebaut, aus ihnen durch Kreuzung neue Arten gezogen und die Ernte aus diesen in einer seinen besonderen Zwecken entsprechenden Weise registriert. Unter anderm wurden die Samenkörner in den einzelnen Schoten gezählt — mit Weglassung tauber Schoten — und die Zahlen der Schoten von einem bis zehn Körnern, der Höchstzahl, die sich überhaupt ergab, festgestellt. So ergab sich die folgende Verteilungstafel der 60 536 geernteten Schoten.

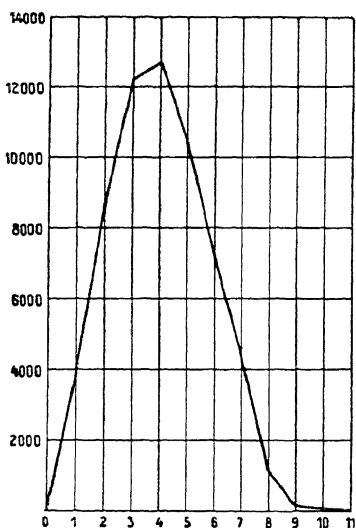



Fig. 7.
Verteilung von Erbsenschoten nach der Zahl der Körner.

Das zugehörige Häufigkeitspolygon, Fig. 7, zeigt deutlich linke Asymmetrie. Der linke Endpunkt O ist feststehend, er würde bleiben, wenn man das Material beliebig vermehrte; hingegen könnte dabei der rechte Endpunkt weiter hinausrücken, wenn sich Schoten mit mehr als zehn Körnern fänden. Einem solchen Häufigkeitspolygon sollte daher eine Kurve unterlegt werden, welche die Achse links schneidet, sich ihr aber rechts unbegrenzt nähert. Indessen steht nichts im Wege, das Polygon einer Kurve anzupassen, die sich beiderseits ins Unendliche erstreckt, nach

Tab. 25. Verteilung der Erbsenschoten nach der Zahl der Körner.

Körner x	Schoten z
1	3 792
2	8 567
3	12 150
4	12 742
5	10 388
6	7 083
7	4 225
8	1 473
9	115
10	1
	60 536

[1]) K. Pearson, Contributions to the Mathematical Theory of Evolution. Phil. Trans. Roy. Soc. of London. A, vol. 186 (1895), p. 364.

[2]) K. Marbe, Die Gleichförmigkeit in der Welt, II. Bd., München 1919, S. 95.

Art von Fig. 6. Da auch nach einer bestimmten Regel, die Bevorzugungen ausschaltet, herausgehobene Teile der Ernte dieselbe Verteilungsform zeigten, so kann diese als für die vorliegende Materie typisch erachtet werden.

Betraf dieses Beispiel ein unstetiges Kollektiv, so beziehen sich die zwei andern auf stetige Kollektive.

2) W. Johannsen[1]) ließ die Längenausdehnung von 558 unbeschädigten Feuerbohnensamen in der Weise bestimmen, daß bei jedem Samen das Millimeterintervall angegeben wurde, in welches die Länge fiel. So ergab sich unmittelbar eine Klasseneinteilung nach Millimetern, die aus der Tab. 26 ersichtlich ist.

Das zugehörige Häufigkeitspolygon, Fig. 8, zeigt neben Unregelmäßigkeiten, die in dem verhältnismäßig kleinen Umfang des Kollektivs und der im Vergleich dazu feinen Klasseneinteilung ihren Grund haben, eine deutliche linksseitige Asymmetrie.

Tab. 26. Verteilung der Längen von Feuerbohnen.

X in mm	z
17—18	3
18—19	7
19—20	21
20—21	23
21—22	53
22—23	69
23—24	85
24—25	75
25—26	72
26—27	56
27—28	39
28—29	25
29—30	21
30—31	4
31—32	4
32—33	1
	558

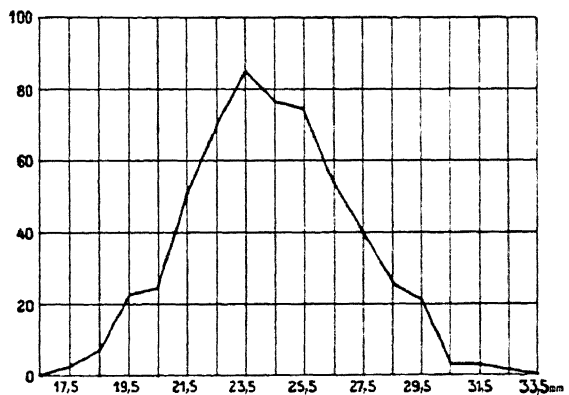



Fig. 8. Häufigkeitspolygon der Länge von Feuerbohnen.

3) Zu Vergleichen geben Anlaß die verwandten Materien, die umstehend in den Tab. 27 und 28 und den dazugehörigen Fig. 9 und 10 vorgeführt sind.

Als bemerkenswert ist die Tatsache einer ausgeprägten Symmetrie bei den Neugeborenen[2]), einer deutlichen linken Asymmetrie bei den Erwachsenen[3]) hervorzuheben. Da sich diese auch nach der Scheidung der Untersuchten nach ihrer engeren Abkunft (Engländer, Schotten, Walen, Irländer) erhalten hat, so dürfte sie eine der Materie eigentümliche Eigenschaft sein.

[1]) W. Johannsen, Elemente der exakten Erblichkeitslehre, 3. Aufl., Jena 1926, S. 13. Vgl. P. Riebesell, Mathematische Statistik und Biometrik, Frankfurt a. M. und Berlin 1932, S. 20.

[2]) Aus der Tafel 15 (Art. 25, 2) durch starke Reduktion entstanden.

[3]) G. U. Yule, An Introduction to the Theory of Statistics, London 1932, p. 95.

Tab. 27. Gewichte von männlichen Neugeborenen.

X in kg	z
1,5—2,0	1
2,0—2,5	14
2,5—3,0	59
3,0—3,5	140
3,5—4,0	53
4,0—4,5	19
4,5—5,0	2
	288

Tab. 28. Gewichte von männlichen Erwachsenen.

X in engl. Pfd.	z	X in engl. Pfd.	z	X in engl. Pfd.	z
90—100	2	160—170	1326	230—240	16
100—110	34	170—180	787	240—250	11
110—120	152	180—190	476	250—260	8
120—130	390	190—200	263	260—270	1
130—140	867	200—210	107	270—280	—
140—150	1623	210—220	85	280—290	1
150—160	1559	220—230	41		7749

Fig. 9. Gewichte von männlichen Neugeborenen.

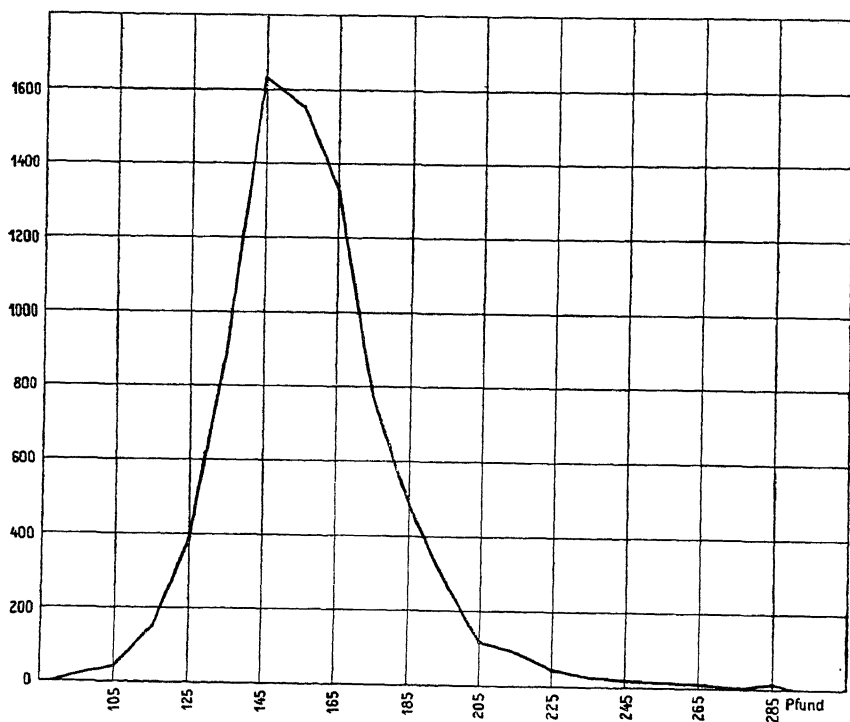
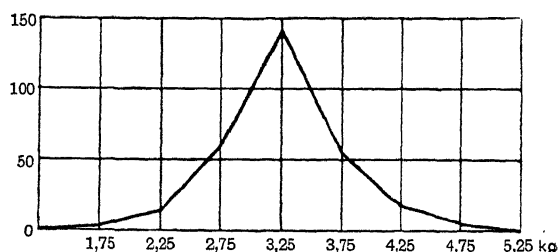



Fig. 10. Gewichte von männlichen Erwachsenen.

32. Als Grenze einer hochgradigen Asymmetrie hat eine Verteilung zu gelten, bei welcher der Abfall der Häufigkeit von einem größten Wert nur nach einer Seite stattfindet, so daß man in Bezug auf den Ausgangspunkt von einer einseitigen Verteilung zu sprechen hat.

Beispiele einer solchen Verteilung bieten sich insbesondere auf wirtschaftlichem Gebiete dar; so kommen die kleinsten Einkommen bei dem größten Teile der Bevölkerung vor und werden mit wachsender Höhe immer weniger häufig; ebenso steht es mit dem Nutzwert der Häuser eines großen Gebietes, die minderwertigen sind in größter Anzahl vorhanden und mit dem steigenden Nutzwert wird die Anzahl immer kleiner.

Tab. 29.

Verteilung der veranlagten Pflichtigen nach Einkommensgruppen im Deutschen Reich 1928.[1])

Klasse	Einkommensgruppen X	Zahl der Pflichtigen z	z durchschnittlich auf ein Intervall von 1500 RM.
1	bis 1 500 RM.	1 864 087	1 864 087
2	über 1 500 „ 3 000 „	1 171 597	1 171 597
3	„ 3 000 „ 5 000 „	504 018	378 013
4	„ 5 000 „ 8 000 „	241 675	120 838
5	„ 8 000 „ 12 000 „	211 461	79 298
6	„ 12 000 „ 16 000 „	86 078	32 279
7	„ 16 000 „ 25 000 „	72 379	12 063
8	„ 25 000 „ 50 000 „	42 276	2 537
9	„ 50 000 „ 100 000 „	12 335	370
10	„ 100 000 „	4 977	

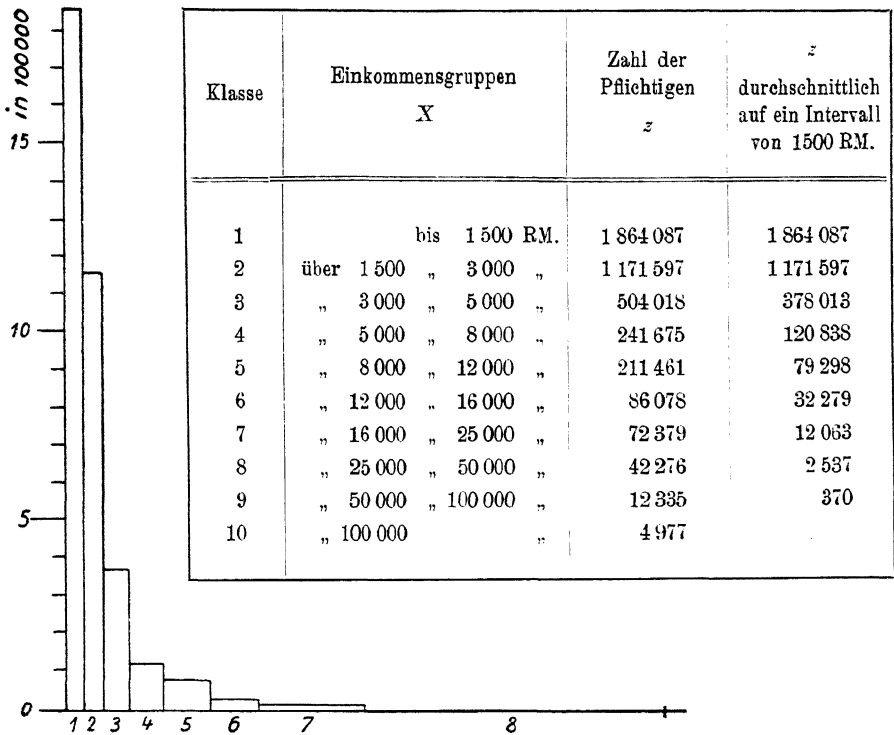



Fig. 11.

Verteilung der veranlagten Pflichtigen nach Einkommensgruppen.

1) Statistik des Deutschen Reichs, Band 391, S. 6.

Tab. 30. Verteilung der Häuser in England und Wales nach dem Nutzwerte 1885 bis 1886.[1])

Klasse	X in £	z	z durchschnittlich auf ein Intervall von 10 £
1	0— 10	3 174 806	3 174 806
2	10— 20	1 450 781	1 450 781
3	20— 30	441 595	441 595
4	30— 40	259 756	259 756
5	40— 50	150 968	150 968
6	50— 60	90 432	90 432
7	60— 80	104 128	52 064
8	80— 100	47 326	23 663
9	100— 150	58 871	11 774
10	150— 300	37 988	2 533
11	300— 500	8 781	439
12	500—1000	3 002	60
13	1000—1500	1 036	21

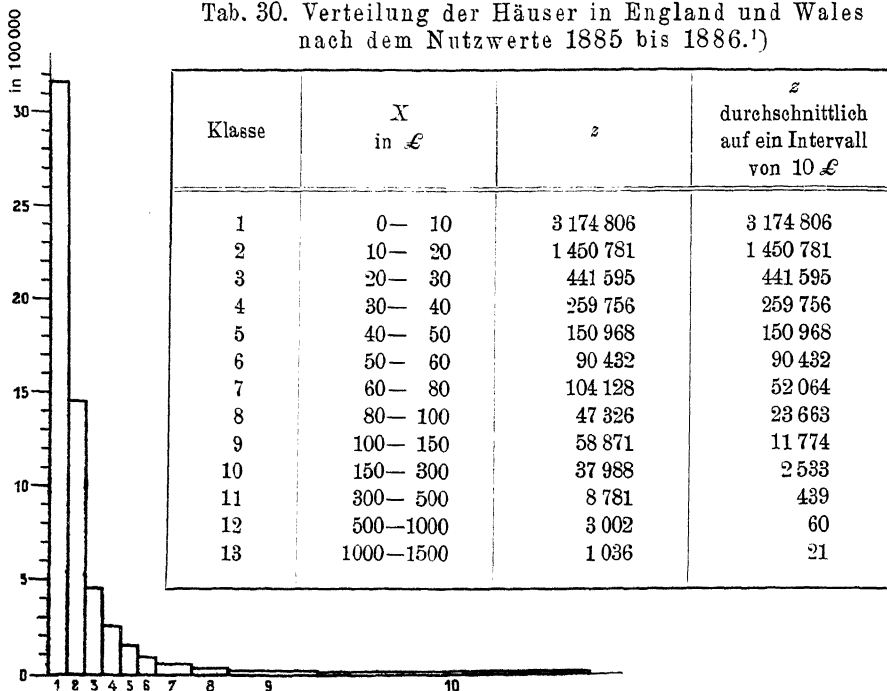



Fig. 12. Verteilung der Häuser nach dem Nutzwert.

Tab. 31. Diphtheriesterbefälle nach Jahrfünften und Jahrzehnten.

X	z
0— 5	49 479
5—10	23 348
10—15	4 092
15—20	1 123
20—25	585
25—35	786
35—45	512
45—55	324
55—65	260
65—75	127
75 u. darüber	35
	80 671

Bei beiden Kollektiven nehmen die Klassen an Größe zu und werden dadurch nicht vergleichbar. Der Wechsel in der Klassengröße zeigt sich beim zweiten Kollektiv an den Unregelmäßigkeiten der z-Kolonne; man achte auf den Übergang von 90 432 zu 104 128, von 47 326 zu 58 871. Um Vergleichbarkeit herzustellen, ist in der vierten Kolonne die durchschnittliche Zahl der auf ein Intervall von 1500 RM. bzw. 10 £ entfallenden Pflichtigen bzw. Häuser angegeben. Nach den Zahlen dieser Kolonne ist Fig. 11 bzw. 12 als Treppenpolygon entworfen. Die eingelegte Häufigkeitskurve hätte die Gestalt eines Hyperbelastes.

Zu einer einseitigen Verteilung führen die schon in Tab. 21 (Art. 26) besprochenen Diphtheriesterbefälle, wenn man die ersten fünf Jahre zusammenzieht, wie das die nebenstehende Tab. 31 und das sie veranschaulichende Diagramm Fig. 13a zur An-

[1]) K. Pearson, Phil. Trans. Roy. Soc. of London, A, vol. 186 (1895), p. 396.

schauung bringen; daneben ist in Fig. 13b das Diagramm zur dritten Kolonne der Verteilungstafel 21 (Art. 26) aufgenommen. Dieses erst läßt erkennen, daß die genannte Krankheit den Höhepunkt ihrer Sterblichkeit im 4. Lebensjahre erreicht und daß es sich in Wirklichkeit um eine zweiseitige, hochgradig asymmetrische Verteilung handelt. Ähnliches würde sich bezüglich der Einkommen und bezüglich der Wohnhäuser ergeben, wenn man die unterste Klasse aufteilen würde; nicht Null ist das häufigste „Einkommen", sondern ein gewisser niedriger Betrag innerhalb der ersten Klasse.

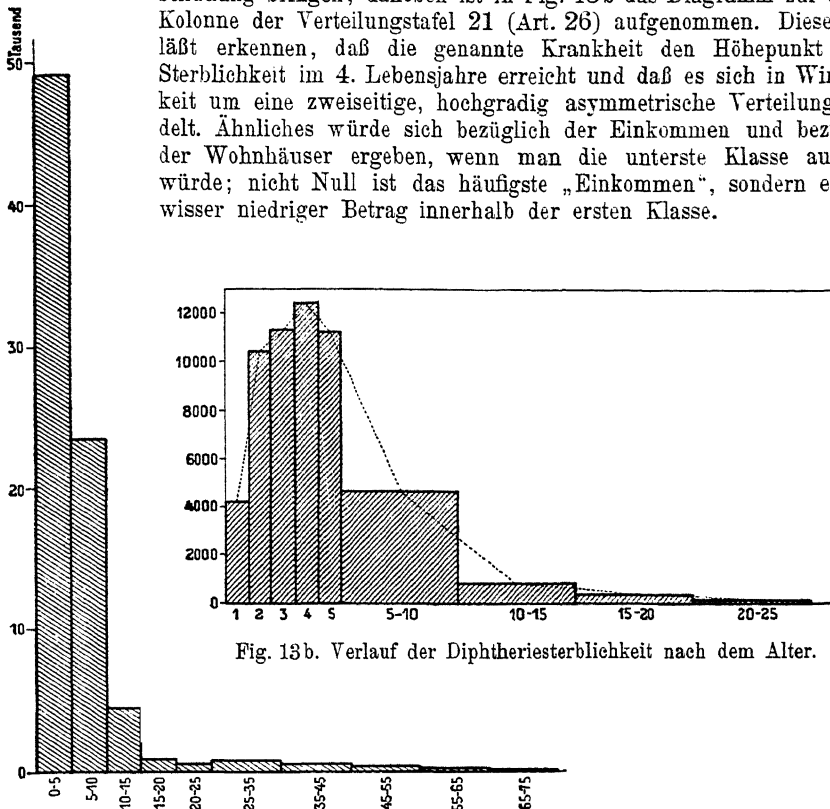



Fig. 13b. Verlauf der Diphtheriesterblichkeit nach dem Alter.

Fig. 13a. Verlauf der Diphtheriesterblichkeit nach dem Alter.

33. Wenn auch die eingipflige, mehr oder weniger asymmetrische bis nahezu symmetrische Verteilung die weitaus am häufigsten anzutreffende ist, so ist sie doch keineswegs die allein herrschende. Es sind auch Verteilungen beobachtet worden, die von dieser wesentlich verschieden sind. Zwei Beispiele solcher besonderen Verteilungen mögen hier vorgeführt werden.

Die Fieder der Esche tragen Blättchen in wechselnder Anzahl, zumeist so, daß sich auch an der Spitze des Fieders ein unpaariges Blättchen findet. Nachstehend ist die Verteilung von 8554 hierauf untersuchten Fiedern angegeben[1]). x in der ersten Zeile ist die Anzahl der Blättchen, z in der zweiten Zeile gibt die Anzahl der Fieder, an welchen jene angetroffen wurde.

x:	3	4	5	6	7	8	9	10	11	12	13	14	15	16
z:	8	5	142	75	876	237	2674	527	2947	223	753	26	59	2.

[1]) K. Pearson, Phil. Trans. Roy. Soc., A, vol. 197 (1901), p. 295.

Die ungeraden Werte von x zeigen ein anderes Verhalten als die geraden[1]; es sind hier zwei Verteilungen übereinandergelagert, jede von ihnen ist angenähert symmetrisch, wie dies noch deutlicher aus der geometrischen Darstellung Fig. 14 zu ersehen ist.

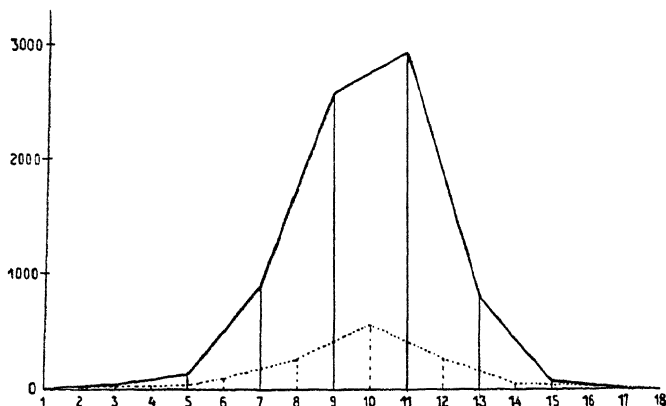



Fig. 14. Verteilung der Eschenfieder nach der Zahl der Blättchen.

Eine der normalen im gewissen Sinne entgegengesetzte Verteilung hat Pearson[2]) an den Graden der Bewölkung beobachtet; statt eines höchsten Punktes zeigt das Häufigkeitspolygon einen tiefsten Punkt, von dem aus es nach beiden Seiten ansteigt gegen die beiden Enden, die einerseits dem vollkommen reinen, anderseits dem völlig bedeckten Himmel entsprechen. Die Beobachtungen, täglich während 10 Jahren in Breslau angestellt, ergaben folgende Reihen: x bedeutet den Grad der Bewölkung, z die Anzahl der Tage, an welchen dieser Grad registriert wurde.

x:	0	1	2	3	4	5	6	7		9	10
z:	751	179	107	69	46	9	21	71	194	117	2089.

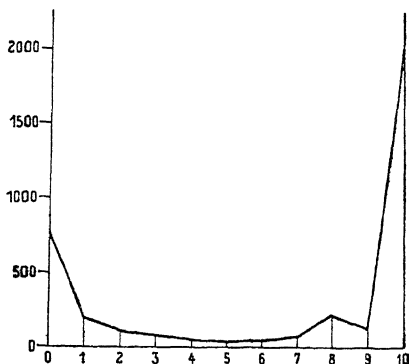



Fig. 15. Verteilung der verschiedenen Grade der Bewölkung.

Die Fig. 15 zeigt das entsprechende Verteilungspolygon, dem eine U-förmige Häufigkeitskurve entsprechen würde.

Es sei erwähnt, daß Pearson[3]) und

---

[1]) Es wurden nur unbeschädigte Fieder gezählt. Über die Entstehung der Fieder mit gerader Blättchenzahl hat Pearson bei dieser Gelegenheit Beobachtungen angestellt. Es findet ein Spaltungsvorgang statt, der Seitenblätter entstehen läßt statt des einen Spitzenblattes.

[2]) Proc. Roy. Soc., A, vol. 62 (1897), p. 287.

[3]) K. Pearson, Contributions to the Mathematical Theory of Evolution Phil. Trans. Roy. Soc. of London, A, vol. 186 (1895), p. 343. Vgl. hierzu P. Riebeseil, Einführung in die Sachversicherungsmathematik, Berlin 1936, S. 27 u. f.

seine Schule es als eine Hauptaufgabe der biologischen und der Variationsstatistik überhaupt angesehen haben, die verschiedenen auftretenden Verteilungen durch analytische Funktionen darzustellen. Für die Bestimmung der Parameter dieser Funktionen aus der beobachteten Verteilung haben sie eine eigene Methode, die Momentenmethode, ausgearbeitet. Es handelt sich im wesentlichen um die Gewinnung empirischer Formeln von angenommenem analytischen Bau auf Grund vorliegender Beobachtungen. — Pearson hat auch den Gedanken verfolgt, daß nichtnormale Verteilungen entstanden sein können durch Übereinanderlagerung von zwei oder mehreren normalen Verteilungen, da bei heterogenen Materien naheliegt, anzunehmen, sie seien durch Mischung zweier oder mehrerer an sich homogenen Materien entstanden. Daraus entsprang die Aufgabe, zu versuchen, eine nichtnormale Verteilung als Summe von zwei oder selbst mehreren normalen Verteilungen darzustellen und die Elemente dieser zu bestimmen. — Selbstverständlich sind alle diese Berechnungen nur so weit durchzuführen, als durch sie neue Erkenntnisse gewonnen werden.

Anschließend sei bemerkt, daß die Häufigkeitskurve für die Anwendung der statistischen Methode in der Technik von großer Wichtigkeit ist. Der Leiter der Forschungsabteilung der Vereinigten Stahlwerke A.-G. in Düsseldorf Dr.-Ing. Karl Daeves[1]) hat in dieser Richtung grundlegende Forschungen angestellt. Daeves untersucht die Häufigkeitskurven für die Zerreißfestigkeit von Stählen mit einem Kohlenstoffgehalt von 0,1 %, 0,2 % und 0,3 %. In der graphischen Darstellung trägt er auf der x-Achse die Festigkeit in kg/mm² und auf der y-Achse die relative Häufigkeit in Prozent der Gesamtzahl der Werte jeder Kurve auf. Die Häufigkeitskurve ist für Stähle mit 0,1 % C linksseitig asymmetrisch und für Stähle mit 0,2 % C und 0,3 % C näherungsweise symmetrisch. Weiter untersucht Daeves die Verteilung der Dehnungsprozente von Kesselblechen mit 0,1 % C, die Verteilung der Schwefelgehalte der Roheisenabstiche bei zwei Hochöfen, die Häufigkeitskurve der Leistung (Lebensdauer) von Werkzeugen.

Die in der Drahtindustrie auftretenden Häufigkeitskurven unterzieht E. Kohlweiler[2]) der Untersuchung.

§ 2. Mittelwerte.

34. Die Beschreibung eines Kollektivs durch Verteilungstafel und Häufigkeitspolygon oder Häufigkeitskurve kann nicht als seine abschließende Erledigung gelten. Es fehlt ihr die Kürze und Bestimmtheit. Man wird sich dessen bewußt, wenn man daran gehen will, zwei Kollektive derselben Art auf dieser Grundlage zu vergleichen; man wird manche Unterschiede in den Einzelheiten durch Worte ausdrücken können, aber eine zusammenfassende Kennzeichnung ihres gegenseitigen Verhältnisses wird sich nicht geben lassen. Eine solche kann nur dadurch geschehen, daß man gewisse, wohl definierte Maße aus dem Beobachtungsmaterial ableitet und diese einander gegenüberstellt.

Worin können sich zwei Kollektive derselben Art der Hauptsache nach voneinander unterscheiden?

---

[1]) K. Daeves, Praktische Großzahl-Forschung. Methoden zur Betriebs-Überwachung und Fehlerbeseitigung. VDI-Verlag, Berlin 1933, S. 25 u. f.

[2]) E. Kohlweiler, Statistik im Dienste der Technik. München und Berlin 1931, S. 124 u. f.

Einmal darin, daß ein geeignet definierter Wert des Arguments bei beiden verschieden groß ausfällt; dann darin, daß sie verschiedene Ausbreitung zeigen, d. h. daß die Argumentwerte bei dem einen sich über einen größeren Teil des Maßstabs verteilen als bei dem andern.

Was den ersten Punkt betrifft, so sind es die verschiedenen Mittelwerte, von welchen man Gebrauch macht.

Unter einem Mittelwert im allgemeinen Sinne versteht man einen irgendwie definierten Wert, der zwischen den kleinsten und größten fällt, die das Kollektiv aufweist. Es läßt sich eine unbegrenzte Zahl mathematischer Konstruktionen ersinnen, die dieser Forderung genügen. Man wird solche bevorzugen, die mit den Eigenschaften der Häufigkeitskurve in einer Beziehung stehen.

Hinsichtlich des zweiten Punktes schiene sich die Angabe der beiden äußersten Argumentwerte, des kleinsten und des größten, darzubieten; aber abgesehen von den später auszuführenden Mängeln, die einem solchen Vorgang anhaften, hat er den Nachteil, aus zwei Zahlen zu bestehen, während die Kennzeichnung durch eine einzige Zahl erwünscht ist. Wir wollen jeder Zahl, welche dies in einer geeigneten Weise leistet, den Namen eines Streuungsmaßes geben.

Damit sind wir zur Theorie der Mittelwerte und der Streuungsmaße geführt, denen indessen nicht bloß ein theoretisches Interesse zukommt, die vielmehr auch bei praktischen Fragen eine entscheidende Rolle spielen können.

35. Es ist schon erwähnt worden, daß sich mathematisch eine unabsehbare Menge von Mittelwerten ausdenken läßt; aber nur solche, die gewissen Forderungen genügen, eignen sich für unseren Zweck; der Grad, in welchem sie diese Forderungen einzeln und in ihrer Gesamtheit erfüllen, entscheidet über ihre Brauchbarkeit.

Eine der ersten Eigenschaften, die ein Mittelwert besitzen soll, ist die leichte Verständlichkeit; Wissenschaft wird nicht um ihrer selbst willen betrieben, ihre Ergebnisse sollen der Allgemeinheit zugänglich sein. Darum sind Mittelwerte von abstrakt mathematischer Definition, mit welcher sich keine anschauliche Vorstellung verbinden läßt, wenig oder gar nicht geeignet.

Zweitens muß verlangt werden, daß sich die Definition praktisch möglichst scharf, unzweideutig verwirklichen läßt. Je weniger dies von der Übung und Geschicklichkeit des Ausführenden abhängt, um so besser.

Man wird drittens fordern, daß sich der Mittelwert auf alle Beobachtungen stütze, so daß er ein Ausdruck ihrer Gesamtheit ist.

Ein vierter Gesichtspunkt wird einem Mittelwert den Vorzug geben, der eine leichte mathematische Handhabung gestattet; wenn z. B. zwei oder mehrere Kollektive, nachdem sie bereits einzeln durchgerechnet sind, zu einem Kollektiv vereinigt werden, so ist es wertvoll, wenn man den Mittelwert des vereinigten Kollektivs aus den bereits vorliegenden Mittelwerten der einzelnen in einfacher Weise ableiten kann.

Es wird fünftens die Leichtigkeit der Rechnung mit in die Waagschale fallen, wenn sie auch den andern Forderungen gegenüber nicht überwiegende Bedeutung haben darf. Dort, wo häufig Mittelwerte zu ziehen sind, fällt möglichste Vereinfachung des Rechnungsverfahrens sehr ins Gewicht; die Ausbildung von Methoden, die leicht und sicher zu handhaben sind, ist darum eine wichtige Angelegenheit.

Es hat sich nur eine kleine Zahl von Mittelwerten eingebürgert, und die Reihenfolge, in der sie hier aufgezählt werden, entspricht ihrer abnehmenden Verbreitung und Bedeutung; es sind dies das arithmetische Mittel, der Zentralwert, der dichteste Wert, das geometrische und das harmonische Mittel[1]).

36. Das arithmetische Mittel. Wenn von einer Größe X eine Anzahl N von besonderen Werten: $X_1, X_2, \ldots X_N$ vorliegt, so nennt man den N-ten Teil ihrer Summe das arithmetische Mittel dieser Werte.

Wird dafür der Buchstabe M eingeführt, so hat man in Zeichen:

$$M = \frac{1}{N} \Sigma\,(X). \tag{1}$$

Die Definition macht keinen Unterschied, ob die Werte alle untereinander verschieden sind oder ob sich darunter Gruppen von gleichen vorfinden. Indem man nur die verschiedenen Werte ansetzt und jedem eine Zahl z beifügt, welche aussagt, wievielmal er sich wiederholt, kommt $N = \Sigma\,(z)$ und die Regel lautet dann:

$$M = \frac{\Sigma\,(z\,X)}{\Sigma\,(z)}. \tag{2}$$

Sie kommt zur Anwendung, wenn das arithmetische Mittel aus einer primären Verteilungstafel berechnet werden soll, während die Formel (1) auf die Urliste anzuwenden wäre.

Die rein arithmetische Definition läßt eine Deutung zu, die eine Beziehung zur Wahrscheinlichkeitsrechnung herstellt, von der mit Nutzen Gebrauch gemacht werden kann.

Gelten die N Werte $X_1, X_2, \ldots X_N$ als gleichberechtigt, so kommt jedem die Wahrscheinlichkeit $\frac{1}{N}$ zu, d. h. mit dieser Wahrscheinlichkeit ist auf Grund der vorliegenden Erfahrung der einzelne Wert bei Wiederholung der Beobachtung zu

[1]) Vgl. hierzu F. Žižek, Grundriß der Statistik, 2. Aufl., München und Leipzig 1923, S. 148 u. f., und Die Statistischen Mittelwerte, Leipzig 1908, und P. Flaskämper, Beitrag zur Logik der statistischen Mittelwerte (Allgemeines Statistisches Archiv, 21. Bd., 1931, S. 379 u. f.). Beide Autoren stellen grundlegende Untersuchungen über den logischen Charakter der Mittelwerte an. Žižek betont die Notwendigkeit der begrifflichen Übereinstimmung der Größen, aus welchen ein Mittelwert berechnet werden soll; er stellt in diesem Zusammenhang ein Postulat der möglichsten Homogenität der Reihen, aus denen Mittelwerte berechnet werden, auf. Flaskämper weist darauf hin, daß die Mittelwerte zu den statistischen Begriffen zu rechnen sind, die auch der Mathematik angehören und die ihre besonderen logischen Eigenschaften haben, welche aus der Natur des Zahlbegriffes sich ergeben. Bei diesen Begriffen kommt zur Logik der Zahlenhaftigkeit noch die Logik der zahlenmäßig zu beschreibenden Situation. Flaskämper gründet auf diese Überlegung sein Postulat vom Parallelismus von Sach- und Zahlenlogik (vgl. auch P. Flaskämper, Die Bedeutung der Zahl für die Sozialwissenschaften. Allgemeines Statistisches Archiv, 23. Bd., 1933. S. 58 u. f.).

erwarten; und kommt der betreffende Wert in der Reihe z mal vor, so ist $\frac{z}{N}$ diese Wahrscheinlichkeit. Nach dieser Auffassung erscheint das arithmetische Mittel als Summe der Produkte der Einzelwerte mit ihren Wahrscheinlichkeiten.

Bei unstetigen Kollektiven, deren Argument eine ganze Zahl ist, ist der Bereich seiner Werte meist ein eng beschränkter, die Rechnung nach der Formel (2) unterliegt dann keiner Schwierigkeit. Das Resultat wird zumeist keine ganze Zahl sein und verlangt eine entsprechende Auslegung.

Bei stetigen Kollektiven großen Umfangs würde sich die Rechnung auf Grund der Primärtafel umständlich gestalten. Man macht dann von einer entsprechend ausgewählten Verteilungstafel Gebrauch und stützt sich dabei auf die Annahme, daß alle Argumentwerte einer Klasse identisch sind mit der Klassenmitte; man ersetzt also gewissermaßen das stetige Kollektiv durch ein unstetiges. Das hat zur Folge, daß man statt des wirklichen arithmetischen Mittels der Einzelwerte nur einen Näherungswert erhält, um so genauer, je kleiner das Klasseninterval und je gleichmäßiger die Einzelwerte über die Intervalle verteilt sind.

Bezeichnet man die Klassenmitte allgemein mit x, so kommt an die Stelle der Formel (2) die der Gestalt nach gleiche, aber dem Sinne nach veränderte Formel

$$M = \frac{\Sigma\,(z\,x)}{\Sigma\,(z)}. \qquad (3)$$

37. Die Ausführung nach dieser Formel würde noch immer eine beträchtliche Arbeit erfordern, wenn die x und z vielziffrige Zahlen sind.

Eine erhebliche Vereinfachung ergibt sich durch folgende zwei Festsetzungen:

1. Als vorläufige Einheit, in welcher die Argumentwerte ausgedrückt werden, wird das Klasseninterval benützt;

2. statt mit den Argumentwerten wird mit ihren Abweichungen ε von einem passend gewählten Ausgangswert gerechnet.

Durch die erste Festsetzung erreicht man, daß die Klassenmitten ausnahmslos, was auch die Klassengröße sein möge, durch ganze Zahlen, die Wechselpunkte durch Zahlen mit dem Anhang 0,5 bezeichnet sind.

Durch die zweite Festsetzung wird die Rechnung auf die möglichst kleinsten Zahlen zurückgeführt, wenn man den Ausgangspunkt beiläufig in die Mitte der Verteilungstafel, und zwar in eine Klassenmitte, verlegt. Heißt er U, so drückt sich eine Klassenmitte allgemein durch $U + \varepsilon$ aus, mithin wird

$$\Sigma\,(z\,x) = U\Sigma\,(z) + \Sigma\,(z\,\varepsilon) \qquad (4)$$

und

$$M = U + \frac{\Sigma\,(z\,\varepsilon)}{\Sigma\,(z)}. \qquad (5)$$

In dem einen Teil der Tafel, dem (räumlich) oberen, sind alle ε negativ, in dem unteren positiv.

Beispiel. Die nachstehende Tabelle zeigt die zweckmäßige Anordnung der Rechnung. Gerechnet wird darin das arithmetische Mittel der Höhen neunjähriger Kiefern aus Verteilungstafel 13 (Art. 25, 1) mit dem Klassenintervall 10 cm.

Berechnung des arithmetischen Mittels.

Klassenmitte x	Häufigkeit z	Abweichung ε	Produkt $z\varepsilon$
60	1	−11	11
70	1	−10	10
80	0	− 9	0
90	0	− 8	0
100	1	− 7	7
110	3	− 6	18
120	3	− 5	15
130	5	− 4	20
140	6	− 3	18
150	11	− 2	22
160	10	− 1	10
			−131
170	17	0	
180	17,5	1	17,5
190	13	2	26
200	13,5	3	40,5
210	6	4	24
220	7	5	35
230	3	6	18
240	2	7	14
250	3	8	24
260	1	9	9
270	1	10	10
	125		+218

$$U = 170; \qquad \Sigma(z\varepsilon) = 218 - 131 = 87; \qquad M = 170 + \frac{87 \cdot 10}{125} = 176{,}96 \text{ cm.}$$

Um den Einfluß der Reduktion und selbst der Reduktionslage auf das arithmetische Mittel zu beleuchten, ist in den nächsten zwei Tabellen die Rechnung auch für die reduzierten Tafeln 14: I, II mit einem Klassenspielraum von 20 cm ausgeführt.

Bestimmung des arithmetischen Mittels aus der reduzierten Tafel.

I.

x	z	ε	$z\varepsilon$
65	2	—5	10
85	0	—4	0
105	4	—3	12
125	8	—2	16
145	17	—1	17
			—55
165	27	0	
185	30,5	1	30,5
205	19,5	2	39
225	10	3	30
245	5	4	20
265	2	5	10
	125		+129,5

$U = 165$; $\Sigma(z\varepsilon) = 129{,}5 - 55 = 74{,}5$;

$$M = 165 + \frac{74{,}5 \cdot 20}{125} = 176{,}9 \text{ cm.}$$

II.

x	z	ε	$z\varepsilon$
55	1	—6	6
75	1	—5	5
95	1	—4	4
115	6	—3	18
135	11	—2	22
155	21	—1	21
			—76
175	34,5	0	
195	26,5	1	26,5
215	13	2	26
235	5	3	15
255	4	4	16
275	1	5	5
	125		+88,5

$U = 175$; $\Sigma(z\varepsilon) = 88{,}5 - 76 = 12{,}5$;

$$M = 175 + \frac{12{,}5 \cdot 20}{125} = 177{,}0 \text{ cm.}$$

Drückt man die mittlere Höhe in Dezimetern aus und behält bloß eine Dezimalstelle bei, was mit der Genauigkeit der Messungen im Einklang stehen dürfte, so erhält man bei den drei Berechnungen für M den Wert 17,7 dm. In diesem Falle ist die Reduktionslage ohne Einfluß auf das arithmetische Mittel.

Es ist nur eine andere Ausbildung des vorstehenden Verfahrens, was Johannsen[1]) als „Umbiegen" der Verteilungsreihe um den Ausgangspunkt bezeichnet. Das Vorzeichen wird dabei von den Abweichungen auf die Häufigkeitszahlen übertragen.

Beispiel. An 703 Butten (Pleuronectes), gefangen in der Umgebung von Skagen, wurden die Strahlen in den Schwanzflossen gezählt. Es wurden gefunden mit

Strahlenanzahl:	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61
Anzahl Butten:	5	2	13	23	58	96	134	127	111	74	37	16	4	2	1.

[1]) W. Johannsen, Elemente der exakten Erblichkeitslehre, 3. Aufl., Jena 1926, S. 11 und 34. Vergl. auch P. Riebesell, Mathematische Statistik und Biometrik, Frankfurt a. M. und Berlin 1932, S. 20.

Wenn 53 als Ausgangswert genommen wird, ergibt die Rechnung:

x		53	54	55	56	57	58	59	60	61	
			52	51	50	49	48	47			(1)
z	+	134	127	111	74	37	16	4	2	1	
	−		96	58	23	13	2	5			(2)
ε		0	1	2	3	4	5	6	7	8	(3)
Diff. (2)	+	134	31	53	51	24	14		2	1	
	−							1			(4)
Prod.	+	0	31	106	153	96	70		14	8	
(3) · (4)	−							6			(5)

$$\text{Summe aus (5): } 478 - 6 = 472; \quad M = 53 + \frac{472}{703} = 53{,}67.$$

Der Sinn dieses Ergebnisses ist der, daß bei möglichst gleichmäßiger Verteilung der Strahlen auf ein Exemplar 53 oder 54 Strahlen entfielen, u. zw. wie eine einfache Rechnung lehrt, auf 232 Exemplare je 53 und auf 471 Exemplare je 54 Strahlen.

38. Mit der jetzt vorzuführenden Berechnungsweise des arithmetischen Mittels legen wir den Grund zu einem Verfahren, das sich im weiteren Verfolge der Untersuchungen als sehr vorteilhaft erweisen wird und wegen des Umstandes, daß dabei nur Additionen vorkommen, Summenverfahren heißen soll. Zur Vorbereitung der erforderlichen Entwicklungen schicken wir das allgemeine Bild der Verteilungstafel in einer dem Zwecke entsprechenden Form voraus, wobei wir uns auf die ersten drei Kolonnen beschränken, die in den vorangehenden speziellen Tabellen vorkamen. Die Mitte der Klasse, welche der ersten besetzten vorausgeht, erhält das Zeichen a; alles übrige ist aus dem nebenstehenden Schema zu ersehen.

Schema der Grundtafel für das Summenverfahren.

Klassenmitte x	Klassenhäufigkeit z	Abweichung vom Ausgangswert U ε
$a+1$	z_1	$-(k-1)$
$a+2$	z_2	$-(k-2)$
$a+3$	z_3	$-(k-3)$
.	.	.
.	.	.
.	.	.
$a+(k-2)$	z_{k-2}	-2
$a+(k-1)$	z_{k-1}	-1
$U=a+k$	z_k	0
$a+(k+1)$	z_{k+1}	1
$a+(k+2)$	z_{k+2}	2
.	.	.
.	.	.
.	.	.
$a+(n-2)$	z_{n-2}	$n-k-2$
$a+(n-1)$	z_{n-1}	$n-k-1$
$a+n$	z_n	$n-k$
	$N=\Sigma(z)$	

Das Verfahren besteht in einem fortschreitenden Aufsummieren der z, das bei langen Tafeln zur Vermeidung großer Zahlen von oben und unten gegen die

Mitte geführt wird, woselbst der Ausgangswert U angenommen wird. Auf diese Art werden folgende Summen gebildet:

$$(6)\begin{cases} s_1 = z_1 \\ s_2 = z_1 + z_2 \\ s_3 = z_1 + z_2 + z_3 \\ \dots\dots\dots\dots \\ s_{k-2} = z_1 + z_2 + z_3 + \dots + z_{k-2} \end{cases} \qquad (7)\begin{cases} s_n = z_n \\ s_{n-1} = z_n + z_{n-1} \\ s_{n-2} = z_n + z_{n-1} + z_{n-2} \\ \cdot \\ s_{k+2} = z_n + z_{n-1} + z_{n-2} + \dots + z_{k+2} \end{cases}$$

$$(8) \qquad s_{k-2} + z_{k-1} = S_0^- \qquad\qquad (9) \qquad s_{k+2} + z_{k+1} = S_0^+$$

Aus (6) und (7) folgt durch Summierung

$$\begin{aligned} s_1 + s_2 + \dots + s_{k-2} &= (k-2)\, z_1 + (k-3)\, z_2 + \dots + z_{k-2} \\ &= (k-1-1)\, z_1 + (k-2-1)\, z_2 + \dots + (2-1)\, z_{k-2} \\ &= -\sum_1^{k-2} (z\,\varepsilon) - s_{k-2} = S_1^-; \end{aligned} \tag{10}$$

$$\begin{aligned} s_n + s_{n-1} + \dots + s_{k+2} &= (n-k-1)\, z_n + (n-k-2)\, z_{n-1} + \dots + z_{k+2} \\ &= (n-k-1)\, z_n + (n-k-1-1)\, z_{n-1} + \dots + (2-1)\, z_{k+2} \\ &= \sum_n^{k+2} (z\,\varepsilon) - s_{k+2} = S_1^+. \end{aligned} \tag{11}$$

Nun ist

$$\begin{aligned} \Sigma (z\,\varepsilon) &= \sum^{k-2} (z\,\varepsilon) - z_{k-1} + z_{k+1} + \sum_{k+2}^{n} (z\,\varepsilon) \\ &= -\left(S_1^- + s_{k-2}\right) - z_{k-1} + z_{k+1} + S_1^+ + s_{k+2} \\ &= S_1^+ - S_1^- + S_0^+ - S_0^- = \Delta_1 + \Delta_0, \end{aligned} \tag{12}$$

wenn

$$\begin{aligned} S_0^+ - S_0^- &= \Delta_0 \\ S_1^+ - S_1^- &= \Delta_1 \end{aligned} \tag{13}$$

gesetzt wird.

Aus

$$\frac{\Sigma (z\,\varepsilon)}{\Sigma (z)} = \eta$$

und dem Ausgangswert U ergibt sich nach (5) das arithmetische Mittel

$$M = U + \eta. \tag{14}$$

Für den vorliegenden Zweck ändern wir das Schema ein wenig ab, indem wir mit der Summenbildung (6) noch um zwei Schritte, bis s_k, und mit der Summenbildung (7) um einen Schritt, bis s_{k+1}, weiter gehen, wir haben dann die Kontrolle, daß

$$s_k + s_{k+1} = \Sigma(z)$$

sein muß; ferner ist dann

$$s_{k-1} = S_0^- \qquad s_{k+1} = S_0^+;$$

es bedarf also nur noch der Bildung von S_1^- und S_1^+.

Beispiele.

1) In diesem Beispiel, das die Höhen der Jungkiefern (Verteilungstafel 13 in Art. 25, 1) betrifft, setzen wir neben die Hauptzahlen noch die entsprechenden Buchstaben der allgemeinen Entwicklung.

Arithmetisches Mittel nach dem Summenverfahren.

x cm	z	s		
60	1	1		
70	1	2		
80	0	2		
90	0	2		
100	1	3		
110	3	6		
120	3	9		
130	5	14		
140	6	20		
150	11	31		
160	10	41	(S_0^-)	90 (S_1^-)
170	17	58		
180	17,5	67	(S_0^+)	151 (S_1^+)
190	13	49,5		
200	13,5	36,5		
210	6	23		
220	7	17		
230	3	10		
240	2	7		
250	3	5		
260	1	2		
270	1	1		
	125			

Kontrolle: $58 + 67 = 125$

$\Delta_0 = 67 - 41 = 26$

$\Delta_1 = 151 - 90 = 61$

$\eta = \frac{26+61}{125} 10 = 7,0$

$M = 170 + 7,0 = 177,0 \text{ cm}$

Bei der Berechnung von η ist jedesmal darauf zu achten, daß es wieder in der ursprünglichen Maßeinheit auszudrücken ist; diese beträgt hier 10 cm, darum die Multiplikation des Bruches, der η in Klassengrößen ausdrücken würde, mit 10.

2) Das zweite Beispiel hat die Körpergewichte männlicher erwachsener Personen (Art. 31, 3, Tab. 28) zum Gegenstande.

Gewichte erwachsener männlicher Personen. Arithmetisches Mittel.

x engl. Pfund	z	s	
95	2	2	
105	34	36	
115	152	188	
125	390	578	
135	867	1445	
145	1623	3068	2249
155	1559	4627	
165	1326	3122	3921
175	787	1796	
185	476	1009	
195	263	533	
205	107	270	
215	85	163	
225	41	78	
235	16	37	
245	11	21	
255	8	10	
265	1	2	
275	—	1	
585	1	1	
	7749		

Probe: $4627 + 3122 = 7749$

$$\Delta_0 = 3122 - 3068 = 54$$

$$\Delta_1 = 3921 - 2249 = 1672$$

$$\eta = \frac{54 + 1672}{7749} \cdot 10 = 2{,}23$$

$$M = 155 + 2{,}23 = 157{,}23 \text{ engl. Pf.}$$

3) Die folgenden zwei Beispiele betreffen unstetige Kollektive, und zwar a) die Zahl der Strahlen in der Schwanzflosse einer Buttenart (Art. 37, zweites Beispiel), b) die Zahl der Samen in 178 Hülsen von Indigofera australis[1]). Die Rechnung ist so angelegt, daß bei a) das niedrigste, bei b) das höchste Argument als Ausgangswert genommen ist; es kommt dann nur der untere, bzw. nur der obere Teil der vorstehenden Tabellen zur Geltung.

[1]) F. Ludwig, Botanisches Zentralblatt, Bd. 73 (1898), S. 348.

Schwanzflossenstrahlen bei Pleuronectes.
Arithmetisches Mittel.

x	z	s	
47	5	703	
48	2	698	3992
49	13	696	
50	23	683	
51	58	660	
52	96	602	
53	134	506	
54	127	372	
55	111	245	
56	74	134	
57	37	60	
58	16	23	
59	4	7	
60	2	3	
61	1	1	
	703		

$$\Delta_0 = 698, \qquad \Delta_1 = 3992$$

$$\eta = \frac{698 + 3992}{703} = 6{,}7$$

$$M = 47 + 6{,}7 = 53{,}7$$

Samenzahl pro Hülse bei Indigofera australis.
Arithmetisches Mittel.

x	z	s	
3	1	1	
4	2	3	
5	8	11	
6	13	24	
7	22	46	
8	45	91	
9	63	154	
10	23	177	330
11	1	178	
	178		

$$\Delta_0 = 0 - 177, \qquad \Delta_1 = 0 - 330$$

$$\eta = -\frac{177 + 330}{178} = -2{,}8$$

$$M = 11 - 2{,}8 = 8{,}2$$

Die Kontrolle der s liegt bei a) in der obersten, bei b) in der untersten Zahl. Über die Deutung solcher nicht ganzzahliger Ergebnisse vergleiche man das im Art. 37 beim zweiten Beispiel Gesagte.

39. Das arithmetische Mittel erfüllt die Forderungen, die an einen Mittelwert gestellt werden, insgesamt und in vorzüglicher Weise; darin findet die hervorragende Stellung, die es einnimmt, und die fast ausschließliche Anwendung ihre wissenschaftliche Begründung.

Der arithmetische Mittel- oder Durchschnittswert[1]) wird häufig im praktischen Leben angewandt (z. B. im Verkehrswesen: Anzahl der in der Zeiteinheit durchschnittlich beförderten Personen, Gütermengen usw.).

An wertvollen arithmetischen Eigenschaften sind die folgenden zu erwähnen:

1. Wird in der Gleichung (5) $U = M$, so muß $\Sigma\,(z\,\varepsilon) = 0$ werden, d. h. die algebraische Summe der Abweichungen vom arithmetischen Mittel ist gleich Null.

[1]) Hinsichtlich der sprachlichen Gleichsetzung von arithmetischem Durchschnitts- und Mittelwert vgl. L. v. Bortkiewicz, Grundriß einer Vorlesung über Allgemeine Theorie der Statistik, Berlin 1912, S. 16.

2. Werden mehrere Kollektive mit den Umfängen N_1, $N_2, \ldots$ zu einem vereinigt, das dann den Umfang $N = N_1 + N_2 + \ldots$ hat, und bezeichnet man die Argumentwerte in den Einzelkollektiven allgemein mit X_1, $X_2, \ldots$, im Gesamtkollektiv mit X, so ist

$$\Sigma X = \Sigma X_1 + \Sigma X_2 + \ldots;$$

formt man diese Gleichung um in

$$N \frac{\Sigma X}{N} = N_1 \frac{\Sigma X_1}{N_1} + N_2 \frac{\Sigma X_2}{N_2} + \ldots$$

so ergibt sich der Satz:

$$NM = N_1 M_1 + N_2 M_2 + \ldots \tag{15}$$

über den Zusammenhang der Mittelwerte. Mit Hilfe von (15) kann man das Gesamtmittel aus den Einzelmitteln ableiten, ohne die umständliche Rechnung einer Mittelbildung nötig zu haben. Wenn ein Kollektiv, für das die Rechnung bereits durchgeführt ist, nachträglich erweitert wird, so braucht man das arithmetische Mittel nur für die Erweiterung zu rechnen und dieses mit dem früheren der Regel (15) gemäß zu verbinden.

3. Ist eine Variable X die Summe oder Differenz zweier andern Variablen X_1, X_2, so ist das arithmetische Mittel von X gleich der Summe, bzw. der Differenz der arithmetischen Mittel der Komponenten.

Liegt für X_1 eine Verteilungstafel vom Umfange N_1, für X_2 eine Verteilungstafel vom Umfange N_2 vor, so kann man durch Zusammenstellung (Addition oder Subtraktion) je eines Wertes von X_1 mit je einem Werte von X_2 einen Wert von X bilden, im Ganzen $N_1 N_2$ Werte; summiert man diese, so ergibt sich

$$\Sigma X = N_2 \Sigma X_1 \pm N_1 \Sigma X_2;$$

durch Division mit $N_1 N_2$ ergibt sich links das arithmetische Mittel M von X, mithin ist tatsächlich

$$M = M_1 \pm M_2. \tag{16}$$

Es ist leicht, diese Entwicklung auf eine algebraische Summe beliebig vieler Variablen auszudehnen.

Hat man beispielsweise Teile eines Ganzen, jeden für sich, gemessen (Glieder von Organen eines Lebewesens) und aus den Messungen eines jeden das Mittel gezogen, so ergibt sich der Mittelwert des Ganzen nach der (erweiterten) Regel (16).

4. Die für das arithmetische Mittel kennzeichnende Beziehung

$$\Sigma (z \varepsilon) = 0$$

gibt an, welche Stellung es im Diagramm einnimmt. Dort sind die z durch Flächenstreifen, insbesondere durch die Rechtecke des Staffelbildes, Fig. 2, die ε durch die Abstände der Mittellinien der Streifen von jener Ordinate dargestellt, deren Fußpunkt das arithmetische Mittel ist; das statische Moment der Diagrammfläche in Bezug auf die letztgenannte Ordinate ist Null, heißt, daß diese Ordinate durch den Schwerpunkt der Fläche geht.

40. Der Begriff des Schwerpunktes hat eine besondere Bedeutung bei zweifach ausgedehnten (zweidimensionalen) statistischen Verteilungen. Darunter versteht man solche Verteilungen, bei denen die Gesamtheit der Elemente eines Kollektivs nach zwei Merkmalen gegliedert wird. Die graphische Darstellung von zweidimensionalen statistischen Verteilungen, die Bestimmung und die Eigenschaften des Schwerpunktes, werden in Art. 75 ausführlich dargelegt.

In der Bevölkerungsstatistik tritt der Schwerpunkt in der Form des Bevölkerungsschwerpunktes[1]) auf. Man denkt sich die Bevölkerung eines Landes auf einer Platte, auf die das geographische Kartenbild übertragen wird, aufgestellt, u. zw. erhält jede Person ihren Platz nach der Lage ihres Wohnortes. Bei sämtlichen Personen wird gleiches Gewicht vorausgesetzt. Es gibt dann einen und nur einen Punkt auf der Platte, in dem diese von unten her unterstützt werden muß, damit sie, ohne an anderen Stellen unterstützt zu werden, in der Gleichgewichtslage bleibt. Diesen Unterstützungspunkt nennt man den Schwerpunkt der Bevölkerung. Die Platte selbst ist hiebei ohne Masse zu denken.

Die Bestimmung des Bevölkerungsschwerpunktes wird in folgender Weise vorgenommen. Es seien für die m Orte eines Gebietes die geographischen Längen bezeichnet mit $l_1, l_2, \ldots l_m$, die geographischen Breiten mit $b_1, b_2, \ldots b_m$ und die Einwohnerzahlen mit $e_1, e_2, \ldots e_m$. Die geographische Länge L und die geographische Breite B des Bevölkerungsschwerpunktes berechnen sich nach der Formel

$$L = \frac{e_1 l_1 + e_2 l_2 + \ldots + e_m l_m}{e_1 + e_2 + \ldots + e_m}$$

$$B = \frac{e_1 b_1 + e_2 b_2 + \ldots + e_m b_m}{e_1 + e_2 + \ldots + e_m}$$

Der Bevölkerungsschwerpunkt des Deutschen Reiches hat sich nach den Feststellungen von G. Wegemann[2]) in der Zeit von 1816—1900 in nordöstlicher Richtung nach der Reichshauptstadt zu verschoben, u. zw. um 10,8 km nach Norden und 13,6 km nach Osten. Von 1900—1910 hat er sich um 4 km nach Osten und um 1,4 km nach Norden, ebenfalls auf Berlin zu, bewegt. Im Jahre 1910 lag er ungefähr 4 km nördlich von Nebra an der Unstrut (Provinz Sachsen). Nach dem neuen Gebietsstand des Reiches befand sich 1910 der Bevölkerungsschwerpunkt 22 km westlich und 1,4 km südlich vom Schwerpunkt 1910 nach dem alten Gebietsstand. Diese Verschiebung zeigt in großer Linie deutlich an, daß der Bevölkerungsverlust im Osten größer war als im Westen. In der Zeit von 1910—1933 hat sich der Schwerpunkt der Bevölkerung des neuen Reichsgebietes um 7 km nach Westen bewegt. Er liegt nach der letzten Volkszählung in der Stadt Heldrungen (Erfurter Becken). Die Verschiebung nach Westen führt zu dem Schluß, daß von 1910—1933 die Bevölkerungsvermehrung im Süden des Deutschen Reiches der im Norden ungefähr die Waage hielt, während, im großen betrachtet, der Bevölkerungszuwachs im Westen wesentlich stärker war als im Osten.

Die Lageveränderung des Bevölkerungsschwerpunktes und des Schwerpunktes im allgemeinen läßt sich übersichtlich mittels der Theorie der komplexen Zahlen

[1]) Vgl. G. v. Mayr, Statistik und Gesellschaftslehre, Bd. II Bevölkerungsstatistik, 2. Aufl., Tübingen 1926, S. 85 und F. Burkhardt, Der statistische Schwerpunkt und seine Bedeutung für Theorie und Praxis. Allgemeines Statistisches Archiv, 19. Bd., 1929, S. 473 u. f.

[2]) Petermanns Mitteilungen, Bd. 49, 1903, S. 210.

darstellen. Zu diesem Ende bringen wir die zweifach ausgedehnte Verteilung in Beziehung zur Zahlenebene. Wir tragen vom Anfangspunkt O des Koordinatensystems aus auf der Abszissenachse (reelle Achse) das erste Merkmal a und auf der Ordinatenachse (imaginäre Achse) das zweite Merkmal b auf. Dem Element mit den Merkmalsvariationen a_k und b_k ordnen wir die komplexe Zahl

$$\zeta_k = a_k + i b_k$$

zu. Der Bildpunkt dieser komplexen Zahl ζ_k in der Zahlenebene sei P_k. Für ein Kollektiv von n Elementen erhalten wir auf diese Weise n komplexe Zahlen:

$$\begin{aligned} \zeta_1 &= a_1 + i b_1 \\ \zeta_2 &= a_2 + i b_2 \\ &\;\vdots \\ \zeta_n &= a_n + i b_n \end{aligned}$$

und ihre n Bildpunkte $P_1, P_2, \ldots P_n$, die auch zum Teil zusammenfallen können.

Den Schwerpunkt P der Verteilungstafel bestimmen wir folgendermaßen: Wir addieren geometrisch die Vektoren $OP_1, OP_2, \ldots OP_n$ und dividieren den so erhaltenen Vektor durch die reelle Zahl n. Der Schwerpunkt P ist der Bildpunkt der komplexen Zahl

$$Z = \frac{\zeta_1 + \zeta_2 + \ldots + \zeta_n}{n} = A + iB,$$

wobei

$$A = \frac{a_1 + a_2 + \ldots + a_n}{n}$$

$$B = \frac{b_1 + b_2 + \ldots + b_n}{n}$$

ist.

Der Schwerpunkt besitzt die für die statistische Praxis wichtige Eigenschaft, daß die Summe der Quadrate der Abstände der einzelnen Bildpunkte vom Schwerpunkt (polares quadratisches Moment) ein Minimum ist. Auf dieser Eigenschaft beruht z. B. die Anwendung von Schwerpunktsberechnungen zur Bestimmung des Standortes von Bahnhöfen, Frauenkliniken u. a.[1])

Mit Hilfe der Theorie der komplexen Zahlen läßt sich die Verschiebung des Bevölkerungsschwerpunktes von einer Volkszählung zur anderen in folgender Weise darstellen. In der Zeit zwischen zwei Volkszählungen vollziehen sich Veränderungen an den Einwohnerzahlen durch die Geburten, Sterbefälle und Wanderungen. Jede Gemeinde hat entweder einen Geburtenüberschuß oder einen Sterbefallüberschuß und weiter entweder einen Zuwanderungs- oder einen Abwanderungsüberschuß. Wir bestimmen die Schwerpunkte für diese Bewegungsmassen, u. zw. zunächst für die Gemeinden mit Geburtenüberschuß den Schwerpunkt für die Mehr-Lebendgeborenen als Gestorbenen. Ebenso bestimmen wir für die Gemeinden mit Sterbefallüberschuß den Schwerpunkt der Mehr-Gestorbenen als Lebendgeborenen, für

[1]) F. Burkhardt, Zur Minimumseigenschaft des arithmetischen Mittels. Deutsches Statistisches Zentralblatt 1926, Sp. 139 u. f.

die Gemeinden mit Zuwanderungsüberschuß den Schwerpunkt der Mehr-Zugewanderten als Abgewanderten und für die Gemeinden mit Abwanderungsüberschuß den Schwerpunkt der Mehr-Abgewanderten als Zugewanderten. Durch die Lage dieser vier Schwerpunkte und die Personenzahlen, auf die sie sich beziehen, ist die Verschiebung des Bevölkerungsschwerpunktes von einer Volkszählung bis zur nächsten vollständig gekennzeichnet.

Um dies klarzulegen und um zu zeigen, wie aus den vier Fortschreibungsschwerpunkten die Lage des neuen Bevölkerungsschwerpunktes resultiert, wollen wir die Einwohnerzahlen der einzelnen Gemeinden nach der alten Volkszählung mit $e_1, e_2, \ldots$ und die entsprechenden Einwohnerzahlen nach der neuen Volkszählung mit $e_1', e_2', \ldots$ bezeichnen. Weiter wollen wir die Gesamteinwohnerzahl nach der alten Zählung mit E und die nach der neuen Zählung mit E' bezeichnen. Außerdem führen wir ein für die Anzahl der Mehr-Lebendgeborenen als Gestorbenen in den einzelnen Gemeinden die Bezeichnung $g_1, g_2, \ldots$, für die Mehr-Gestorbenen als Lebendgeborenen die Bezeichnung $h_1, h_2, \ldots$, für die Mehr-Abgewanderten als Zugewanderten die Bezeichnung $u_1, u_2, \ldots$ und für die Mehr-Zugewanderten als Abgewanderten die Bezeichnung $v_1, v_2, \ldots$ Es ist ohne weiteres klar, daß in einer Gemeinde mit Geburtenüberschuß der g-Wert eine positive Zahl und der h-Wert gleich Null ist. In einer Gemeinde mit Sterbefallüberschuß ist es umgekehrt. Das gleiche gilt von den Wanderungswerten. Die Summenzahl für die Mehr-Lebendgeborenen sei G, für die Mehr-Gestorbenen H, für die Mehr-Abgewanderten U und für die Mehr-Zugewanderten V. Wir wollen nun weiter die Lage des Bevölkerungsschwerpunktes nach der alten Zählung mit Z_e und die Lage nach der neuen Zählung mit Z_e' bezeichnen. Schließlich sei noch die Lage der einzelnen Gemeinden durch $\zeta_1, \zeta_2, \ldots$, die Lage des Schwerpunktes der Mehr-Lebendgeborenen durch Z_g, der Mehr-Gestorbenen durch Z_h, der Mehr-Abgewanderten durch Z_u und der Mehr-Zugewanderten durch Z_v gekennzeichnet.

Wir gehen nun so vor, daß wir für die Lage des alten Bevölkerungsschwerpunktes die Beziehung

$$Z_e = \frac{e_1 \zeta_1 + e_2 \zeta_2 + \ldots}{E}$$

und für die Lage des neuen die Beziehung

$$Z_e' = \frac{e_1' \zeta_1 + e_2' \zeta_2 + \ldots}{E'}$$

ansetzen. Die Addition, Multiplikation und Division ist dabei nach den Regeln über das Rechnen mit komplexen Zahlen vorzunehmen. Die neuen Einwohnerzahlen gehen aus den alten in folgender Weise hervor:

$$e_1' = e_1 + g_1 - h_1 - u_1 + v_1$$
$$e_2' = e_2 + g_2 - h_2 - u_2 + v_2 \quad \text{usw.}$$

Setzen wir nun in der Gleichung für den neuen Bevölkerungsschwerpunkt die Ausdrücke für die neuen Einwohnerzahlen ein, so erhalten wir

$$Z_e' = \frac{e_1 \zeta_1 + e_2 \zeta_2 + \ldots}{E'} + \frac{g_1 \zeta_1 + g_2 \zeta_2 + \ldots}{E'} - \frac{h_1 \zeta_1 + h_2 \zeta_2 + \ldots}{E'}$$
$$- \frac{u_1 \zeta_1 + u_2 \zeta_2 + \ldots}{E'} + \frac{v_1 \zeta_1 + v_2 \zeta_2 + \ldots}{E'}.$$

Es läßt sich unter Benützung der vereinbarten Abkürzungen für den neuen Bevölkerungsschwerpunkt die folgende Lagebeziehung aufstellen:

$$Z'_e = \frac{E}{E'} Z_e + \frac{G}{E'} Z_g - \frac{H}{E'} Z_h - \frac{U}{E'} Z_u + \frac{V}{E'} Z_v.$$

Es ist also möglich, die Verschiebung des Bevölkerungsschwerpunktes aus den Schwerpunkten der Mehr-Lebendgeborenen, der Mehr-Gestorbenen, der Mehr-Zugewanderten und der Mehr-Abgewanderten auf graphischem Wege zu ermitteln. Wir reduzieren zunächst den Radius Vector des alten Bevölkerungsschwerpunktes im Verhältnis der alten Einwohnerzahl zur neuen. Weiter reduzieren wir den Radius Vector des Schwerpunktes der Mehr-Lebendgeborenen im Verhältnis ihrer Zahl zur neuen Einwohnerzahl. In entsprechender Weise verfahren wir mit den übrigen drei Schwerpunkten. Die so erhaltenen Punkte setzen wir nach den Regeln über das Arbeiten mit komplexen Zahlen zusammen und kommen auf diese Weise zu dem neuen Bevölkerungsschwerpunkt.

41. Der Zentralwert, auch Medianwert, ist jener Wert des Arguments, der den Umfang des geordneten Kollektivs in zwei gleiche Teile teilt. Argumentwerte, die unter dem Zentralwert liegen, sind also zusammen ebenso häufig wie die Argumentwerte über ihm. Er werde mit C bezeichnet.

Denkt man sich die Argumentwerte eines Kollektivs steigend geordnet, gleiche so oft angesetzt, wie sie vorkommen, so ist der mittelste von ihnen der Zentralwert. Einen mittelsten Wert gibt es aber nur bei ungerader Gliederzahl, $2n+1$; der $(n+1)$te (vom Anfange oder vom Ende) ist dann C. Bei gerader Gliederzahl, $2n$, gibt es zwei mittelste Werte, den n-ten vom Anfang und den n-ten vom Ende; jeder Argumentwert, der zwischen ihnen liegt, sofern sie verschieden sind, erfüllt diese Definition, somit bleibt C innerhalb der bezeichneten Grenzen unbestimmt. Man kann die Definition für diesen Fall dadurch ergänzen, daß man festsetzt, es sei die Mitte zwischen den genannten zwei Punkten für C zu nehmen.

Bei einem unstetigen Kollektiv gibt es streng genommen nur dann einen Zentralwert, wenn sein Umfang N eine ungerade Zahl $2n+1$ ist, wenn man bei fortschreitendem Summieren der Häufigkeiten z zu der Zahl n gelangt und wenn der nächstfolgende Argumentwert mit der Häufigkeit 1 auftritt; denn nur dann gibt es ebensoviel Glieder unter diesem Wert wie über ihm. Da so spezielle Bedingungen kaum jemals erfüllt sein werden, so ist der Begriff des Zentralwerts in seiner strengen Fassung auf unstetige Kollektive gar nicht anwendbar.

Die Bestimmung des Zentralwerts bei einem stetigen Kollektiv geschieht in der Weise, daß man vom Beginn oder vom Ende der Verteilungstafel bis zu jenem Wechselpunkt fortschreitet, bis zu welchem die Summe der Häufigkeiten entweder gleich $\frac{N}{2}$ ist oder möglichst nahe unter dieser Zahl liegt; findet Gleichheit statt, so bezeichnet der betreffende Wechselpunkt selbst den Zentralwert: im andern Falle fällt dieser in das nächste Klassenintervall, und zwar so, daß er es in demselben Verhältnis teilt, in welchem sich der Fehlbetrag auf $\frac{N}{2}$ zur Häufigkeit dieser Klasse befindet. Eine nach dem Summenverfahren angelegte Tafel leistet dabei gute Dienste.

Das Summenverfahren verwendet F. Savorgnan[1]) in graphischer Behandlung zur Bestimmung des Zentralwertes. Er trägt in einem Koordinatensystem (x = Alter, y = Bevölkerungszahl) zu jedem x die Zahl der Person auf, die jünger als x Jahre sind und geht auf der Ordinatenachse vom Anfangspunkt bis zur Hälfte der Gesamtbevölkerungszahl nach oben. Auf diese Weise gelangt er zu einem bestimmten Punkt der Kurve. Der zu diesem Kurvenpunkte gehörige x-Wert ist gleich dem Zentralwert C (Medianalter). Die eine Hälfte der Bevölkerung ist jünger, die andere älter als C Jahre.

Savorgnan bestimmt auch das arithmetische Mittel (Durchschnittsalter) auf graphischem Wege, indem er in dem gleichen Koordinatensystem zu jedem x die Zahl der Personen aufträgt, die älter als x Jahre sind. Das Flächenstück, das von den beiden Koordinatenachsen und der Kurve begrenzt wird, stellt die Summe aller Lebensjahre der Bevölkerung dar. Durch Division des Flächeninhalts durch die Gesamtbevölkerung erhält er das Durchschnittsalter. Savorgnan berechnet auf diese Weise folgende Werte:

	Median-alter	Durchschnitts-alter
Deutsches Reich 1933	30,4	32,6
Frankreich 1931	31,9	34,0
Großbritannien 1931	30,1	32,5
Italien 1931	25,7	29,6

Beispiele. Die nachstehend erläuterten Fälle knüpfen an die in Art. 38 vorgeführten Verteilungen an.

1) Nach der Tabelle über die Strahlenzahl der Schwanzflosse von Pleuronectes ist 703 der Umfang; hiernach hätten unter und über dem Zentralwert je 351 Exemplare zu liegen; es liegen aber über der Strahlenzahl 54, die am ehesten in Betracht käme, deren 245, unter ihr 331; ein Zentralwert im Sinne der Definition besteht also nicht.

2) Die Tabelle über die Samen in den Hülsen von Indigofera australis weist einen Umfang von 178 auf, die Hälfte davon ist 89; am besten entspräche noch das Argument 8, doch liegen nur 46 Fälle darunter (87 darüber), der Definition ist also weitaus nicht Genüge geleistet.

3) Bei den Höhen von Jungkiefern stellt sich die Berechnung von C wie folgt.

Bis 175 cm beträgt die Summe der z 58, der halbe Umfang ist 62,5; demnach ist der Eingriff des Zentralwerts in die nächste Klasse, d. i. die Klasse 175 bis 185,

$$\frac{62,5 - 58}{17,5} \cdot 10 = 2,57,$$

folglich ist

$$C = 175 + 2,57 = 177,57 \text{ cm}.$$

4) In der Tabelle betreffend die Körpergewichte erwachsener männlicher Personen beträgt der halbe Umfang 3874,5; die Summe bis zum Wechsel-

[1]) F. Savorgnan, Die Alterskurve der Bevölkerung, graphische Darstellung der Altersgliederung. Deutsches Statistisches Zentralblatt 1936, Sp. 129 u. f.

punkt 150 ist 3068, der Eingriff in die nächste Klasse, d. i. die Klasse 150—160, beträgt also

$$\frac{3874{,}5 - 3068}{1559} \cdot 10 = 5{,}17,$$

daher hat man

$$C = 150 + 5{,}17 = 155{,}17 \text{ engl. Pfund.}$$

Vom untern Ende der Tafel gelangt man zum Zentralwert durch folgende Schlußweise: bis zum Wechselpunkt 160 beträgt die Summe 3122; der Eingriff in das Intervall 160—150 berechnet sich zu

$$\frac{3874{,}5 - 3122}{1559} \cdot 10 = 4{,}83,$$

somit ist

$$C = 160 - 4{,}83 = 155{,}17 \text{ engl. Pfund.}$$

42. Der Zentralwert erfüllt die allgemeinen Forderungen an einen Mittelwert nur in beschränktem Maße. Leichte Verständlichkeit kommt ihm zu und dürfte der Hauptgrund seiner Einführung gewesen sein; auch bietet seine Bestimmung keine Schwierigkeit. Wiewohl er sich auf alle Beobachtungen stützt, kommt in ihm doch nicht ihre eigentliche Größe, sondern nur ihre Größenanordnung zum Ausdruck; man könnte die unter ihm liegenden Werte beliebig vermindern und die über ihm liegenden beliebig vergrößern, ohne daß er selbst eine Änderung erführe. Diese Unempfindlichkeit schmälert seinen Erkenntniswert. Von den algebraischen Vorteilen, durch die sich das arithmetische Mittel auszeichnet, kommt ihm keiner zu; wenn man die Zentralwerte zweier Kollektive ermittelt hat, so lassen sie keinen Schluß zu auf den Zentralwert, der ihrer Vereinigung zukommt; ebensowenig hängt der Zentralwert der Summe oder Differenz zweier Variablen in irgend einer allgemein ausdrückbaren Weise von den Zentralwerten der einzelnen Variablen ab. Der am schwersten wiegende Nachteil liegt aber in der mangelnden Strenge seiner Bestimmung, die sich auf die meist wenig zutreffende Vorstellung der gleichmäßigen Ausbreitung der Kollektivglieder über das Intervall stützt, in welchem der Zentralwert zu suchen ist.

Die Beziehung des Zentralwertes zum Häufigkeitsdiagramm ist offenkundig; er ist der Fußpunkt jener Ordinate, welche seine Fläche halbiert. Bei vollkommener Symmetrie fällt also der Zentralwert mit dem arithmetischen Mittel zusammen, weicht hingegen um so mehr von ihm ab, je mehr die Verteilung sich von der Symmetrie entfernt. Seine Mitbestimmung hat also Wert für die Beurteilung der Asymmetrie.

Unter gewissen Voraussetzungen[1]) läßt sich auf die Größenbeziehung zwischen

[1]) Es wird angenommen, die Asymmetrie sei so beschaffen, daß gleiche Ordinatenpaare links und rechts auf der einen Seite durchwegs näher beieinanderliegen als auf der andern, so daß die Kurve in entsprechenden Punkten auf der einen Seite durchwegs steiler ist als auf der andern, wobei als entsprechend Punkte bezeichnet werden, die durch Parallele zur Grundlinie ausgeschnitten werden. Man vergleiche hierzu H. E. Timerding, Die Analyse des Zufalls, in „Die Wissenschaft, Einzeldarstellungen aus der Naturwissenschaft und der Technik", Bd. 56, Braunschweig 1915, S. 80.

arithmetischem Mittel und Zentralwert ein Schluß ziehen. In Fig. 16 sei C der Zentralwert; dann halbiert die durch ihn gehende Ordinate die Fläche. Sind σ', σ'' die Schwerpunkte der linken und rechten Hälfte, so halbiert der Schwerpunkt Σ

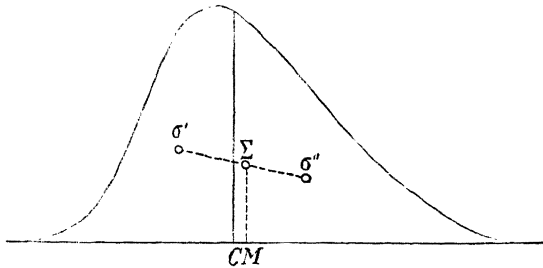



Fig. 16. Größenbeziehung zwischen arithmetischem Mittel und Zentralwert.

der ganzen Fläche die Strecke $\sigma' \sigma''$; Σ aber bestimmt in seiner Abszisse das arithmetische Mittel M. Besteht linke Asymmetrie, so liegt σ' näher an der C-Ordinate als σ'', infolgedessen liegt Σ rechts von ihr, d. h. es ist $M > C$. Bei rechtsseitiger Asymmetrie wird $M < C$.

43. Der dichteste Wert, das Dichtemittel, von den englischen Statistikern Mode genannt. Bei der Betrachtung einer Verteilungstafel sucht das Auge unwillkürlich nach jener Stelle, wo die Häufigkeit am größten ist, also nach jenem Argumentwert, der in dem Kollektiv am häufigsten vertreten ist.

Bei einem unstetigen Kollektiv, wo die Häufigkeitszahl anzeigt, wie oft der betreffende Argumentwert in aller Strenge vorkommt, ist jener Argumentwert, bei dem die größte Häufigkeitszahl steht — vorausgesetzt, daß nur eine solche vorkommt — auch schon der dichteste Wert.

Bei einem stetigen, in Klassen eingeteilten Kollektiv ist durch die größte Häufigkeitszahl vorerst nur die Klasse bezeichnet, in der man den dichtesten Wert zu suchen hat. Man hält sich nämlich an die Vorstellung, daß es eine Häufungsstelle gibt, die mit wachsender Gliederzahl immer deutlicher hervortreten würde, und diese Häufungsstelle bezeichnet den dichtesten Wert. Bei der Klasseneinteilung fällt sie notwendig in die am stärksten besetzte Klasse, wenn sie nicht zufällig mit einer Klassengrenze zusammenfällt.

Der einfachste Fall, den wir zu behandeln haben werden, ist der, daß nur eine Klasse mit größter Häufigkeitszahl vorkommt. Er bildet die Regel bei Kollektiven größeren Umfangs. In der primären Verteilungstafel, die noch alle erhobenen Argumentwerte mit ihren Häufigkeiten aufführt, gibt sich der dichteste Wert in der Regel noch nicht zu erkennen; hier findet gewöhnlich ein Auf- und Abschwanken der Häufigkeitszahlen statt, und eine und dieselbe Zahl, auch die größte, kann wiederholt erscheinen. Das ändert sich meist mit der Klasseneinteilung, und wenn es auch da nicht geschieht, so wird es durch eine geeignete Reduktion der Verteilungstafel herbeigeführt. Indessen kann es auch im Wesen des Kollektivs liegen, daß

nicht ein Maximum der Häufigkeit, sondern deren zwei oder selbst mehr vorhanden sind; das wird z. B. dann der Fall sein, wenn das Kollektiv nicht homogen ist, wenn seine Glieder vielmehr zwei oder noch mehr Kollektiven entstammen, die sich in der Verteilung und insbesondere in ihren dichtesten Werten unterscheiden, wodurch dann in der Vereinigung zwei oder selbst mehrere Gipfel hervortreten. Eine zu weit gehende Reduktion kann die sehr wesentliche Erscheinung verwischen.

Wenn in einer Verteilungstafel, sei es die primäre, kein deutlicher Hinweis auf einen dichtesten Wert zu erkennen ist, so hat man, statt an eine Zusammenziehung der Tafel zu schreiten, einen andern Weg versucht, um zum Ziele zu kommen. Man hat zu jedem Argumentwert statt der ihm eigentümlichen Häufigkeitszahl die Summe oder das arithmetische Mittel aus ihr selbst und den ν vorausgehenden und den ν nachfolgenden, im ganzen also aus $2\nu+1$ Häufigkeitszahlen gesetzt, wobei am Anfang und am Ende leere Klassen zu Hilfe genommen werden müssen. Es ist dies nichts anderes als ein Verfahren zur Ausgleichung der Schwankungen, die man für sachwidrig hält, und zwar ist es, wenn man arithmetische Mittel gebraucht, das von Th. Wittstein angegebene Ausgleichungsverfahren für statistische Tafeln. Abgesehen davon, daß auch dieser Vorgang, selbst wenn man mit ν über 1, 2 hinausgeht, den angestrebten Zweck nicht immer erreicht, ist es nicht ratsam, von der einmal gegebenen Erfahrungsgrundlage zu einer künstlich abgeänderten überzugehen.

Gesetzt nun, es sei eine Klasse mit größter Häufigkeitszahl vorhanden, dann entsteht noch die Frage, an welche Stelle innerhalb derselben der dichteste Wert zu verlegen ist. Darauf haben nicht bloß die benachbarten Klassen, darauf hat die ganze Verteilung Einfluß. Die richtige Antwort könnte nur die ideelle Häufigkeitskurve geben, die man sich als Grenzkurve der beobachteten Verteilung zu denken hat. In Ermangelung ihrer Kenntnis sollte man eine Kurve von bestimmtem analytischem Bau zugrunde legen und der beobachteten Verteilung so gut als möglich anpassen. Pearson hat hierfür zu einer Anzahl praktisch vorkommender Verteilungsformen Methoden ausgearbeitet. Ist einmal die Gleichung der Kurve hergestellt, dann handelt es sich nur noch um die Bestimmung des Maximums der Ordinate; die zugehörige Abszisse ist der gesuchte dichteste Wert. Aber einer und derselben Verteilung lassen sich verschiedene Kurventypen anpassen und auch die Forderung „nach dem besten Anschluß“ ist keine eindeutig zu erfüllende Aufgabe.

Aus diesen Erwägungen geht hervor, daß dem dichtesten Werte neben dem Vorzuge leichter Verständlichkeit der Nachteil anhaftet, daß es nicht möglich ist, für ihn eine eindeutige Bestimmungsweise anzugeben.

Bevor wir zu angenäherten Bestimmungen übergehen, soll über die Bedeutung des dichtesten Wertes einiges gesagt werden. In sachlicher Hinsicht ist sie darin zu erblicken, daß es jener Argumentwert ist, um den sich die Glieder des Kollektivs am dichtesten scharen, so daß man bei der zugrunde liegenden Materie von einer Tendenz sprechen kann, gerade diesen Wert vorzugsweise hervorzubringen. Daraus erklärt sich auch die Wahl der englischen Benennung „Mode“, womit die Art und Weise gemeint ist, wie etwas ist oder geschieht, das Vorbild für etwas; man könnte auch sagen, im dichtesten Wert sei der Typus der im Kollektiv vereinigten Gegenstände zu erblicken. In formaler Hinsicht bewirkt der dichteste Wert eine natürliche Zweiteilung der Verteilung

in einen linken aufsteigenden und einen rechten abfallenden Ast. Darum spielt der dichteste Wert in Fechners „Kollektivmaßlehre“[1]) eine so hervorragende Rolle, weil er der Ausgangswert ist, von dem aus die beiden Äste der Verteilungskurve getrennt behandelt werden. Dem entspricht auch der große Aufwand an Arbeit und Mühe, den Fechner auf sich genommen hat, um zu einer möglichst scharfen Bestimmung des dichtesten Wertes zu gelangen. Die beiden Bestimmungsweisen, die er ausgearbeitet hat, die „empirische“ und die „nach dem Proportionalitätsgesetz“, geben mitunter erheblich abweichende Resultate. Rechtfertigen läßt sich die in vielen Fällen weit getriebene Schärfe der Rechnung nicht. Auf die Frage, welcher der beiden Bestimmungen der Vorrang einzuräumen sei, ließe sich keine begründete Antwort geben.

Wir behandeln nachstehend zwei Fälle: Die Verteilungskurve zeigt 1. nur ein Maximum, 2. mehrere Maxima. In letzterem Falle könnte die Schwierigkeit zumeist durch eine Änderung der Klasseneinteilung behoben werden; es kann aber Gründe geben, die für ein Verbleiben bei der gewählten Einteilung sprechen.

44. Zu einer genäherten, in vielen Fällen ausreichend genauen Bestimmung des Dichtemittels, dem wir das Zeichen D geben, führt die Annahme, die Häufigkeitskurve könne im Bereich der drei Klassen, deren mittlere die am stärksten besetzte ist, durch eine Parabel der Gleichungsform

$$y = \alpha x^2 + \beta x + \gamma$$

ersetzt werden. Als Ursprung werde die Mitte des mittleren Intervalls, als Einheit die Klassengröße genommen; dann haben die vier in Betracht kommenden Wechselpunkte die Abszissen $-\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}$, Fig. 17; und bezeichnet man die drei Häufig-

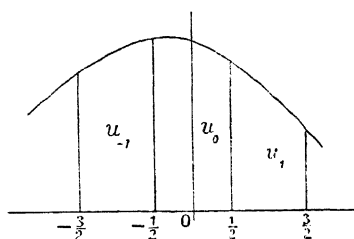



Fig. 17. Zur Bestimmung des dichtesten Wertes.

keiten in der Ordnung von links nach rechts mit u_{-1}, u_0, u_1, so sind sie durch die über den Klassenintervallen ruhenden Flächenstreifen dargestellt. Man hat also zur Bestimmung der Koeffizienten die Gleichungen:

$$u_{-1} = \int_{-\frac{3}{2}}^{-\frac{1}{2}} y\,dx = \frac{13\alpha}{12} - \beta + \gamma,$$

[1]) G. Th. Fechner, Kollektivmaßlehre, Leipzig 1897, S. 182 u. f.

$$u_0 = \int_{\frac{1}{2}}^{\frac{1}{2}} y\,dx = \frac{\alpha}{12} \qquad + \gamma,$$

$$u_1 = \int_{\frac{1}{2}}^{\frac{3}{2}} y\,dx = \frac{13\,\alpha}{12} + \beta + \gamma;$$

durch zweimalige Differenzenbildung entsteht daraus

$$\begin{aligned} \Delta\, u_{-1} &= \quad u_0 - \quad u_{-1} = -\alpha + \beta, \\ \Delta\, u_0 \; &= \quad u_1 - \quad u_0 \;\; = \quad \alpha + \beta, \\ \Delta^2 u_{-1} &= \Delta u_0 - \Delta u_{-1} = \quad 2\alpha. \end{aligned}$$

Nun gibt die Bedingung für das Maximum von y

$$0 = 2\alpha x + \beta$$

$$x = -\frac{\beta}{2\alpha},$$

und zählt man den Abstand des Dichtemittels vom untern Wechselpunkt der stärkstbesetzten Klasse aus, so beträgt er

$$x + \frac{1}{2} = \frac{\alpha - \beta}{2\alpha} = -\frac{\Delta u_{-1}}{\Delta^2 u_{-1}}, \tag{17}$$

wohlbeachtet, in Klasseneinheiten; der Übergang zur ursprünglichen Einheit geschieht durch Multiplikation mit der Klassengröße k. Ist X der zum Wechselpunkt $-\frac{1}{2}$ gehörige Argumentwert, so hat man

$$D = X - k \cdot \frac{\Delta u_{-1}}{\Delta^2 u_{-1}}. \tag{18}$$

Beispiele. 1) Für die wiederholt benützten Jungkiefern (Art. 25, 1), haben wir auf Grund der Verteilungstafel 13 mit der Klassengröße 10 cm die folgende Rechnung:

$$\begin{array}{llll} u_{-1} = 17 & \Delta u_{-1} = \quad 0{,}5 & \Delta^2 u_{-1} = -5{,}0 & -\frac{\Delta u_{-1}}{\Delta^2 u_{-1}} \cdot 10 = 1 \\ u_0 = 17{,}5 & \Delta u_0 = -4{,}5 & & \\ u_1 = 13 & & & \end{array}$$

$$D = 175 + 1 = 176 \text{ cm}.$$

2) Die nachstehende Verteilungstafel betrifft die Rekrutenmaße von 2047 zwanzigjährigen Leipziger Studenten, erhoben in den Jahren 1843 bis 1862 [1]). Die Maßeinheit ist der sächsische Zoll = 23,6 mm. Entwickelt sind alle drei bisher besprochenen Mittelwerte nach den begründeten Regeln.

[1]) G. Th. Fechner, Kollektivmaßlehre, Leipzig 1897, S. 28, 136 und 137.

956,5 1376,5
·767 −1037
189,5 + 339,5
529 : 2047 = 0,26

$M = 71{,}5 + 0{,}26 = 71{,}76$

1023,5
− 767
256,5 : 323,5 = 0,79

$C = 71 + 0{,}79 = 71{,}79$

271 52,5 − 71,0
323,5 − 18,5
305 52,5 : 71,0 = 0,74

$D = 71 + 0{,}74 = 71{,}74$

Die drei Mittelwerte gehören einer und derselben Klasse an und liegen so nahe aneinander, daß man den kleinen Differenzen keine reale Bedeutung beilegen und von einer symmetrischen Verteilung sprechen kann.

Tab. 32. Die Mittelwerte M, C, D von Leipziger Studenten.

x	z	s	
59,5	0,5	0,5	
60,5	0,5	1	
61,5	0	1	
62,5	0	1	
63,5	1	2	
64,5	8	10	
65,5	20	30	
66,5	41,5	71,5	
67,5	72	143,5	
68,5	137	280,5	
69,5	215,5	496	
70,5	271	767	1037
71,5	323,5	1090,5	
72,5	305	956,5	1376,5
73,5	274,5	651,5	
74,5	183,5	377	
75,5	101,5	193,5	
76,5	52	92	
77,5	27,5	40	
78,5	7	12,5	
79,5	3	5,5	
80,5	1,5	2,5	
81,5	0	1	
82,5	1	1	
	2047		

3) Zur Illustration der Bemerkungen, die in Artikel 43 über die Aufsuchung des dichtesten Wertes vorausgeschickt worden sind, bietet die Verteilungstafel 33 des Körpergewichts von 4469 Leipziger Schulmädchen im Alter von 13—14 Jahren geeigneten Stoff. Sie zeigt die Eigentümlichkeit von drei getrennten Maxima in den Häufigkeiten, und in der Annahme, daß es sich vielleicht doch nur um eine zufällig gestörte eingipflige Verteilung handelt, wird man versuchen, ihr durch eine Ausgleichung der Schwankungen näher zu kommen.

Vorstehend sind neben die Häufigkeiten z deren „Summen zu 3", daneben auch die „arithmetischen Mittel zu 3" gesetzt. In der Tat scheint dadurch das Ziel erreicht, denn es tritt nur noch ein Maximum der Häufigkeit auf. Zu seiner Bestimmung können ebensowohl die Summen wie die arithmetischen Mittel verwendet werden; letztere sind als das Ergebnis der Ausgleichung hierhergesetzt und um darauf aufmerksam zu machen, daß ihre Summe stets hinter dem Umfang des Kollektivs zurückbleibt, weil die Randzahlen nicht so oft in Rechnung gezogen sind wie die übrigen.

Aus der letzten Kolonne rechnet man:

415,3	39,4	− 50,4
454,7	− 11,0	
443,7		

$$D = 34,0 + \frac{39,4}{50,4} \cdot 1,5 = 35,17\ kg.$$

Die unveränderte Tafel gibt nach unserem Verfahren als dichtesten Wert 36,12.

Tab. 33. Verteilung der Leipziger Schulmädchen im Alter von 13—14 Jahren nach dem Körpergewicht im Jahre 1922.[1])

Gewicht in kg X	Zahl z	Summen zu 3	Arithm. Mittel zu 3
20,5—22,0	2	4	1,3
22,0—23,5	2	28	9,3
23,5—25,0	24	82	27,3
25,0—26,5	56	176	58,7
26,5—28,0	96	323	107,7
28,0—29,5	171	548	182,7
29,5—31,0	281	806	268,7
31,0—32,5	354	1090	363,3
32,5—34,0	455	1246	415,3
34,0—35,5	437	1364	454,7
35,5—37,0	472	1331	443,7
37,0—38,5	422	1330	443,3
38,5—40,0	436	1202	400,7
40,0—41,5	344	1063	354,3
41,5—43,0	283	863	287,7
43,0—44,5	236	688	229,3
44,5—46,0	169	520	173,3
46,0—47,5	115	398	132,7
47,5—49,0	114	229	76,3
	4469		4430,3

Tab. 34. Verteilung der Leipziger Schulmädchen im Alter von 13—13½ und 13½—14 Jahren nach dem Körpergewicht im Jahre 1922.

Gewicht in kg X	13 bis 13½ z	Gewicht in kg X	13½ bis 14 z
20,0—21,5	1	21,0—22,5	1
21,5—23,0	0	22,5—24,0	4
23,0—24,5	8	24,0—25,5	12
24,5—26,0	30	25,5—27,0	14
26,0—27,5	54	27,0—28,5	46
27,5—29,0	89	28,5—30,0	72
29,0—30,5	169	30,0—31,5	109
30,5—32,0	204	31,5—33,0	153
32,0—33,5	275	33,0—34,5	152
33,5—35,0	276	34,5—36,0	203
35,0—36,5	277	36,0—37,5	172
36,5—38,0	254	37,5—39,0	199
38,0—39,5	218	39,0—40,5	202
39,5—41,0	199	40,5—42,0	154
41,0—42,5	155	42,0—43,5	141
42,5—44,0	115	43,5—45,0	124
44,0—45,5	71	45,0—46,5	83
45,5—47,0	69	46,5—48,0	62
47,0—48,5	53	48,0—49,5	33
48,5—50,0	16		1936
	2533		

[1]) Größe und Gewicht der Schulkinder und andere Grundlagen für die Ernährungsfürsorge, Berlin 1924, S. 34 u. 35. Weiteres wertvolles Zahlenmaterial über Körpergröße und Körpergewicht von Schulkindern hat Prinzing in seinem Handbuch der medizinischen Statistik, 2. Aufl., Jena 1931, S. 121 u. f. zusammengetragen. Es sei auch auf die Untersuchungen von W. Koch, Über die Veränderung menschlichen Wachstums im ersten Drittel des 20. Jahrhunderts, Leipzig 1935, hingewiesen. Koch stellt eine Beschleunigung und Verkürzung des Wachstumsablaufs fest. Vgl. in diesem Zusammenhang auch die Abhandlungen von v. Brunn, Gewichte und Größe der Schulkinder 1920—1932 (Med. Welt 1933, 23), und Risel, Längenmaße und Gewichte der Leipziger Schulkinder (Zeitschr. f. Gesundh. Verwaltung 1933, Bd. 4, S. 145).

Teilt man das Altersjahr 13—14 in die Halbjahre 13—$13^1/_2$ und $13^1/_2$—14 auf, so ergeben sich aus Tabelle 34 für die beiden Halbjahre die dichtesten Werte 35,06 und 35,43. Zwischen diesen beiden Werten liegt das unter Verwendung der „arithmetischen Mittel zu 3“ berechnete Dichtemittel, während das nach der unveränderten Tafel bestimmte Dichtemittel über den beiden Teilmitteln liegt. Es ist also in diesem Beispiel der Berechnung von D mit Hilfe der „arithmetischen Mittel zu 3“ der Vorzug zu geben. Jedoch muß immer von Fall zu Fall entschieden werden, welche der beiden Methoden am Platze ist.

45. Eine andere Behandlung erfordert der Fall, bei dem zwei benachbarte Klassen größte und gleiche Besetzung zeigen. Man wird bei dieser Sachlage die Häufigkeitskurve im Bereich von vier Klassen, den beiden erwähnten und den beiderseits sich anschließenden, durch eine Parabel dritter Ordnung

$$y = \alpha x^3 + \beta x^2 + \gamma x + \delta$$

ersetzen. An der Hand der Fig. 18 ergibt sich folgende Rechnung:

$$u_{-2} = \int_{-2}^{-1} y\,dx = -\frac{15}{4}\alpha + \frac{7}{3}\beta - \frac{3}{2}\gamma + \delta$$

$$u_{-1} = \int_{-1}^{0} y\,dx = -\frac{1}{4}\alpha + \frac{1}{3}\beta - \frac{1}{2}\gamma + \delta$$

$$u_{1} = \int_{0}^{1} y\,dx = \frac{1}{4}\alpha + \frac{1}{3}\beta + \frac{1}{2}\gamma + \delta$$

$$u_{2} = \int_{1}^{2} y\,dx = \frac{15}{4}\alpha + \frac{7}{3}\beta + \frac{3}{2}\gamma + \delta.$$

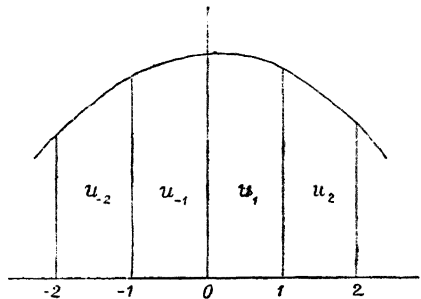



Fig. 18. Zur Bestimmung des dichtesten Wertes.

Zur Bestimmung des Maximums von y hat man die Bedingungsgleichung

$$0 = 3\alpha x^2 + 2\beta x + \gamma,$$

aus der

$$x = \frac{-\beta \pm \sqrt{\beta^2 - 3\alpha\gamma}}{3\alpha} \tag{19}$$

folgt. Aus den obigen vier Gleichungen findet man

$$\begin{aligned} \alpha &= \frac{u_2 - u_{-2}}{6} - \frac{u_1 - u_{-1}}{2} \\ \beta &= \frac{u_2 + u_{-2}}{4} - \frac{u_1 + u_{-1}}{4} \\ \gamma &= \frac{5(u_1 - u_{-1})}{4} - \frac{u_2 - u_{-2}}{12} \end{aligned} \tag{20}$$

und auf Grund der Voraussetzung $u_{-1} = u_1$ vereinfachen sich α und γ:

$$\alpha = \frac{u_2 - u_{-2}}{6}$$

$$\gamma = -\frac{u_2 - u_{-2}}{12}.$$

Welcher von den beiden Werten für x nach (19) Geltung hat, wird sich zumeist aus der Erwägung entscheiden lassen, daß D notwendig im Bereich der zwei inneren Intervalle zu suchen ist.

Beispiel. Einen Fall von der hier betrachteten Art bietet die erste der zwei folgenden Verteilungstafeln, die sich beide auf die Vertikalumfänge von 450 europäischen Männerschädeln beziehen, die Fechner[1]) auf Grund der Messungen Welckers zusammengestellt hat.

Tab. 35. Vertikalumfänge europäischer Männerschädel.

a)

X in mm	z
365,5—370,5	1
370,5—375,5	2
375,5—380,5	5
380,5—385,5	17
385,5 390,5	24
390,5—395,5	36
395,5—400,5	41
400,5—405,5	59
405,5—410,5	65
410,5—415,5	65
415,5—420,5	51
420,5—425,5	40
425,5—430,5	17
430,5—435,5	19
435,5—440,5	4
440,5—445,5	2
445,5—450,5	2
	450

b)

X in mm	z
367,5—372,5	3
372.5—377,5	1
377,5—382,5	7
382,5—387,5	22
387,5—392,5	30
392,5—397,5	33
397,5—402,5	55
402,5—407,5	50
407,5—412,5	73
412,5—417,5	52
417,5—422,5	55
422,5—427,5	35
427,5—432,5	12
432,5—437,5	14
437,5—442,5	5
442,5—447,5	2
447,5—452,5	1
	450

[1]) G. Th. Fechner, Kollektivmaßlehre, Leipzig 1897, S. 102.

Aus $u_{-2} = 59$, $u_{-1} = 65$, $u_1 = 65$, $u_2 = 51$ berechnen sich

$$\alpha = -\frac{4}{3}, \quad \beta = -5, \quad \gamma = \frac{2}{3},$$

$$x = -\frac{5 \pm \sqrt{27\frac{2}{3}}}{4} = \begin{cases} -2{,}57 \text{ Klassenintervalle} \\ 0{,}065 \quad \text{„} \qquad = 0{,}33 \text{ mm}; \end{cases}$$

von den Lösungen hat die zweite Geltung, denn die erste führt über den Bereich der zwei stärkstbesetzten Klassen hinaus. Demnach ist

$$D = 410{,}5 + 0{,}33 = 410{,}83 \text{ mm}.$$

Wenn nicht besondere Gründe für die Tafel a) sprechen, kann man der aufgetretenen Erscheinung durch Abänderung der Klasseneinteilung ausweichen. So weist die Tafel b) nur ein Maximum auf und führt zu folgender Rechnung:

50	23	−44
73	−21	
52		

$$D = 407{,}5 + \frac{23}{44} \cdot 5 = 410{,}11 \text{ mm}.$$

Die beiden Verteilungen geben also nicht unbeträchtlich verschiedene Bestimmungen für D. Das liegt übrigens auch an dem verhältnismäßig kleinen Umfang des Kollektivs.

Die folgende Probe soll zeigen, daß bei größerem Umfang die Bestimmung des dichtesten Wertes viel weniger von der Klasseneinteilung beeinflußt wird. Aus vier verschiedenen Verteilungen von 2047 Leipziger Studentenrekruten [1]) sind jeweils die für die Berechnung von D maßgebenden drei Klassen herausgehoben, daran schließt sich die Rechnung:

I.

Zoll	z		
70 —71	271	52,5	−71,0
71 —72	323,5	−18,5	
72 —73	305		

$$D = 71 + \frac{52{,}5}{71} = 71{,}74$$

II.

Zoll	z		
70,25—71,25	280	47	−70
71,25—72,25	327	−23	
72,25—73,25	304		

$$D = 71{,}25 + \frac{47}{70} = 71{,}92$$

III.

Zoll	z		
70,5—71,5	290	40,5	−75,0
71,5—72,5	330,5	−34,5	
72,5—73,5	296		

$$D = 71{,}5 + \frac{40{,}5}{75} = 72{,}04$$

IV.

Zoll	z		
70,75—71,75	309	9	−41,5
71,75—72,75	318	−32,5	
72,75—73,75	285,5		

$$D = 71{,}75 + \frac{9}{41{,}5} = 71{,}97.$$

Die größte Spannung beträgt bloß $72{,}04 - 71{,}74 = 0{,}30$ Zoll.

[1]) G. Th. Fechner, Kollektivmaßlehre, S. 136 u. 137.

46. Über die Anordnung der drei bisher erledigten Mittelwerte bei asymmetrischer Verteilung läßt sich in Ergänzung dessen, was in Art. 42 bezüglich des arithmetischen Mittels und des Zentralwerts ausgesagt wurde, noch folgendes feststellen. Wie ein Blick auf nebenstehende Fig. 19 lehrt, kommt bei linksseitiger Asymmetrie der Gipfelpunkt in die linke Hälfte der Fläche, folglich ist $C > D$, und da nach früherem in diesem Falle $M > C$ ist, so hat man im ganzen folgende Größenfolge:

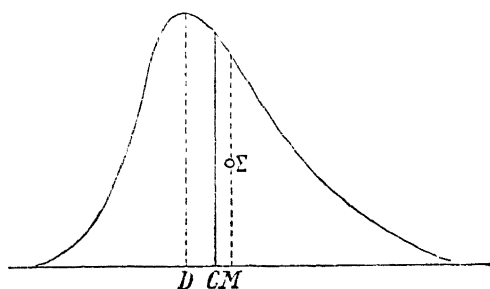



Fig. 19. Lagenbeziehung von M, C, D.

$$D < C < M.$$

Sie kehrt sich um bei rechtsseitiger Asymmetrie in $M < C < D$.

Beispiele: 1) Die nachstehend ausgewiesene Verteilung von 8689 in einem englischen Spital zur Aufnahme gelangten Fällen von Typhoidfieber [1]) (Enteric fever) zeigt ausgesprochene linksseitige Asymmetrie und eignet sich daher zur Illustration des ersten Falles.

Tab. 36. Auftreten von Typhoidfieber nach dem Alter.

X in Jahren	z	s	
0—5	266	266	
5—10	1143	1409	266
10—15	2019	3428	
15—20	1955	5261	7656
20—25	1319	3306	
25—30	857	1987	
30—35	503	1130	
35—40	299	627	
40—45	163	328	
45—50	98	165	
50—55	40	67	
55—60	14	27	
60—65	8	13	
65—70	4	5	
70—75	1	1	
	8689		

$$\begin{array}{rl} 5261 & 7656 \\ -1409 & -\ 266 \\ \hline 3852 & +7390 = 11\,242 \end{array}$$

$$M = 12{,}5 + \frac{11242}{8689} \cdot 5 = 18{,}97$$

$$4344{,}5 - 3428 = 916{,}5$$

$$C = 15 + \frac{916{,}5}{1955} \cdot 5 = 17{,}34$$

$$\begin{array}{lll} 1143 & 876 & -940 \\ 2019 & -\ 64 & \\ 1955 & & \end{array}$$

$$D = 10 + \frac{876}{940} \cdot 5 = 14{,}66.$$

In der Tat ist $D < C < M$ und die Unterschiede sind recht erheblich.

[1]) K. Pearson, Contributions to the Mathematical Theory of Evolution. Phil. Trans. Roy. Soc. of London, A, vol. 186 (1895), p. 390.

2) Die Sterbefälle an Zuckerkrankheit weisen eine rechtsseitig asymmetrische Verteilung auf.

$$M = 85 - \frac{5327 + 9213}{5426} \cdot 10 = 58{,}2$$

$$C = 60 + \frac{2713 - 2332}{1967} \cdot 10 = 61{,}9$$

1206 761 −1700
1967 −939
1028

$$D = 60 + \frac{761}{1700} \cdot 10 = 64{,}5.$$

Wie vorauszusehen war, ist jetzt $M < C < D$.

Tab. 37. Sterbefälle an Zuckerkrankheit im Jahre 1928 in Preußen.[1])

X in Jahren	z	s	
0—10	72	72	
10—20	153	225	
20—30	236	461	
30—40	237	698	
40—50	428	1126	
50—60	1206	2332	
60—70	1967	4299	
70—80	1028	5327	9213
80—90	99	5426	
	5426		

Über die bloße Anordnung von M, C, D hinaus läßt sich noch einiges sagen. Die Unterschiede $C - M$, $M - D$, $D - C$ hängen von der ganzen Verteilung ab, insbesondere von dem Grade der Asymmetrie. Aber auch ihre Verhältnisse, die die gegenseitige Lage kennzeichnen, sind Funktionen der ganzen Verteilung und daher von Fall zu Fall andere; sie unterliegen jedoch nicht so großen Schwankungen wie die absoluten Werte und halten sich bei Verteilungen, die auch nur ähnlich verlaufen, ziemlich konstant. Fechner hat über diese Lagenverhältnisse eingehende Untersuchungen angestellt, die sich auf die Annahme stützen, eine asymmetrische Häufigkeitskurve lasse sich aus zwei halben Normalkurven zusammensetzen; indessen kommt den von ihm so genannten Lagengesetzen keine erhebliche praktische Bedeutung zu.

Als eine Art Faustregel könnte die hingestellt werden, daß bei Verteilungen, die nicht einen übermäßigen Grad von Asymmetrie zeigen, der Zentralwert den Abstand zwischen arithmetischem Mittel und dichtestem Wert im Verhältnis: 1 : 2 teilt, so daß $|M - C|$ etwa $^1/_3\,|M - D|$ beträgt[2]). Man kann dies als eine Art Probe benützen; eine erhebliche Abweichung von diesem Sachverhalt weist auf eine Verteilung hin, die sich der „normalen" auch im erweiterten Sinne der Asymmetrie nicht unterordnen läßt.

In den nachstehenden drei Beispielen wird das Verhältnis der beiden Differenzen $|M - C|$ und $|M - D|$ berechnet.

[1]) Medizinalstatistische Nachrichten, herausgegeben vom Preußischen Statistischen Landesamt, 17. Jahrgang, 2. Heft, Berlin 1930, S. 166.

[2]) Dieses Verhältnis hat Pearson unter Zugrundelegung einer bestimmten Häufigkeitskurve gefunden, nämlich der Kurve $y = y_0\, x^p\, e^{-\gamma x}$, bezogen auf das arithmetische Mittel als Anfangspunkt der Abszissen. Phil. Trans. Roy. Soc. of London, A, vol. 186 (1895), p. 375.

1. Eine später (Art. 60) vorzuführende Tafel der Körpergrößen amerikanischer Rekruten ergibt

$$M = 66{,}701, \qquad C = 66{,}651, \qquad D = 66{,}578 \text{ Zoll.}$$

Die Größenfolge entspricht linksseitiger Asymmetrie; weiter ist

$$M - C = 0{,}050, \qquad M - D = 0{,}123,$$

$^1/_3\,(M - D)$ beträgt 0,041; hier fällt $M - C$ zwischen $^1/_2$ und $^1/_3$ von $M - D$, $(M - C):(M - D) = 1:2{,}46$.

2. Aus der ersten Tafel dieses Artikels — Typhoidfieber — ergab sich

$$M = 18{,}97, \qquad C = 17{,}34, \qquad D = 14{,}66 \text{ Jahre,}$$

die Größenfolge entspricht wieder linksseitiger Asymmetrie; $M - C = 1{,}63$, verglichen mit $M - D = 4{,}31$, zeigt das Verhältnis $1:2{,}64$ gegenüber $1:3$.

3. Die zweite Tafel dieses Artikels — Sterbefälle an Zuckerkrankheit — führte zu

$$M = 58{,}2, \qquad C = 61{,}9, \qquad D = 64{,}5 \text{ Jahre.}$$

Es ist $(C - M):(D - M) = 3{,}7:6{,}3 = 1:1{,}7$.

47. Das geometrische Mittel. Unter dem geometrischen Mittel von n beobachteten Werten $X_1, X_2, \ldots X_n$ einer Variablen X versteht man die (absolute) n-te Wurzel aus ihrem Produkt. Wird also für dieses Mittel der Buchstabe G verwendet, so lautet die Definition in Zeichen:

$$G = (X_1\, X_2 \ldots X_n)^{\frac{1}{n}} \tag{21}$$

Während die bisher betrachteten Mittelwerte unbeschränkte Geltung hatten, müssen hier Null und negative Werte ausgeschlossen werden; denn wenn auch nur einer der Werte Null wird, wird es auch G, und bei negativen X_i kann G reelle Bedeutung verlieren.

Von diesen Ausnahmen abgesehen, ist die Definition völlig bestimmt. Die Berechnung von G wird bei Werten von n, die 3 übersteigen, naturgemäß auf logarithmischem Wege geschehen, der auch bei $n = 2$, 3 im allgemeinen der vorteilhaftere ist. Dadurch kommt das geometrische Mittel in einen Zusammenhang mit dem arithmetischen, indem der Logarithmus des geometrischen Mittels gleich ist dem arithmetischen Mittel der Logarithmen der einzelnen Werte:

$$\log G = \frac{1}{n}(\log X_1 + \log X_2 + \ldots + \log X_n). \tag{22}$$

Mit dem arithmetischen Mittel hat das geometrische manche algebraische Vorteile gemein, die eine Folge dieses Zusammenhangs sind; so besteht zwischen dem geometrischen Mittel eines zusammengesetzten Kollektivs und den geometrischen Mitteln seiner Komponenten eine Beziehung, die nichts anderes ist als der Ausdruck der analogen Beziehung bei arithmetischen Mitteln, nämlich (Art. 39 (15))

$$N \log G = N_1 \log G_1 + N_2 \log G_2 + \ldots \tag{23}$$

Ist die Verteilung der X_i in Bezug auf ihr arithmetisches Mittel symmetrisch, so gilt nicht das Gleiche für die Verteilung der $\log X_i$ in Bezug auf $\log G$.

Einer symmetrischen Verteilung der $\log X_i$ um $\log G$ würde eine eigenartige Verteilung der X_i selbst um G entsprechen. Zwei Werte von $\log X$, die von $\log G$ gleich weit, aber in entgegengesetzter Richtung entfernt sind, so daß

$$\log X' - \log G = -(\log X'' - \log G),$$

entsprechen reziproke Werte des Verhältnisses $\frac{X}{G}$; ist $\frac{X'}{G} = q$, so ist $\frac{X''}{G} = \frac{1}{q}$; das gäbe also eine zur G-Ordinate asymmetrische Verteilung von solcher Art, daß die Abszissen zu gleichen Ordinaten einerseits und anderseits reziprok sind; G erscheint dabei als Dichtemittel.

Daß das geometrische Mittel eine nennenswerte Verwendung nicht gefunden hat, liegt in erster Linie daran, daß es eine abstrakte mathematische Größe ist, die einer anschaulichen Bedeutung ermangelt. Dazu kommt als zweiter Grund die größere Umständlichkeit der Berechnung.

Ein natürlicher Anlaß zur Verwendung des geometrischen Mittels ergäbe sich, wenn man aus irgend einem Grunde die Variation der Glieder eines Kollektivs, statt nach den Differenzen von einem Mittelwert, nach den Verhältnissen zu einem solchen beurteilen wollte. Welcher Mittelwert würde sich dazu besonders eignen? Bezeichnet man ihn vorweg mit G, so hat man es mit der Verteilung der Quotienten

$$\frac{X_1}{G}, \frac{X_2}{G}, \ldots \frac{X_n}{G}$$

zu tun. Ihre Logarithmen sind zum Teil positive, zum Teil negative Zahlen δ_1, $\delta_2, \ldots \delta_n$, weil die Verhältnisse selbst teils über, teils unter 1 liegen, wenn G einen „Mittelwert" bedeuten soll,

$$\log X_1 - \log G = \delta_1, \ \log X_2 - \log G = \delta_2, \ldots \log X_n - \log G = \delta_n;$$

wählt man G so, daß $\Sigma\delta = 0$, eine gewissermaßen indifferente Wahl, so kommt

$$\log X + \log X_2 + \ldots + \log X_n - n \log G = 0,$$

woraus sich

$$G = (X_1 X_2 \ldots X_n)^{\frac{1}{n}}$$

ergibt.

Eine solche Auffassung läßt sich damit rechtfertigen, daß die Variationen in einem Kollektiv mit der Größe seiner Glieder zusammenhängen, daß z. B. die Variationen an einem kleinen Lebewesen sich innerhalb viel engerer Grenzen abspielen als bei einem großen, daß es daher eine gewisse Berechtigung hat, statt Differenzen Verhältnisse zu bilden und sich so von der absoluten Größe der Kollektivglieder unabhängig zu machen. Aber auch dieser Gesichtspunkt hat nicht durchzugreifen vermocht gegenüber den großen Vorzügen des arithmetischen Mittels und selbst gegenüber den beiden anderen Mittelwerten, denen wenigstens der Vorzug leichter Erfaßbarkeit zukommt.

Spielt so das geometrische Mittel bei der Untersuchung von Kollektiven eine untergeordnete Rolle, so tritt es auch sonst nur wenig in die Erscheinung. Ein Fall, wo es praktische Verwendung finden könnte, kommt in der Bevölkerungslehre vor. Wenn man annehmen darf, daß eine Bevölkerung von Jahr zu Jahr nach einer geometrischen Progression wächst, dann ergibt sich die Bevölkerung in der Mitte einer Periode als geometrisches Mittel ihrer Stände am Anfang und am Ende. Bezeichnet V die Größe der Bevölkerung allgemein, V_0 ihren Stand am Anfang, V_n ihren Stand am Ende der n Jahre umfassenden Periode, r den Wachstumsfaktor, so ist

$$V_n = V_0 \, r^n$$

$$V_0 \, V_n = V_0^2 \, r^n$$

$$\sqrt{V_0 \, V_n} = V_0 \, r^{\frac{n}{2}} = V_{\frac{n}{2}}.$$

Aber die Vorausetzungen dieser Ableitung sind solcher Art, daß sie kaum jemals und besonders nicht für lange Zeiträume erfüllt sein werden; das geometrische Mittel kann also im besten Falle als genäherte Schätzung gelten.

Im Deutschen Reich betrug die Einwohnerzahl nach der Volkszählung am 1. Dezember 1890: 49 428 470 und nach der Volkszählung am 1. Dezember 1900: 56 367 178 [1]).

Das arithmetische Mittel dieser beiden Volkszählungszahlen stellt sich auf 52 897 824 und das geometrische Mittel auf 52 783 900. Bei der Volkszählung am 2. Dezember 1895 wurden 52 279 901 [1]) Einwohner gezählt. Das geometrische Mittel liegt somit etwas näher an der letzteren Volkszählungszahl als das arithmetische Mittel. Hieraus folgt, daß die Anwendung des geometrischen Mittels im Zeitraum 1890—1900 sachgemäßer ist als die des arithmetischen Mittels. Der Grund hierfür liegt darin, daß in dem Zeitraum von 1890—1900 die Reichsbevölkerung stark anwuchs. Bei raschem Bevölkerungswachstum vermehrt sich eine Bevölkerung in den einzelnen Teilzeiträumen relativ gleichmäßig, so daß der Wachstumsfaktor in den einzelnen Teilzeiträumen nahezu konstant ist. Bei vollkommener Konstanz des Wachstumsfaktors würde die nach der Methode des geometrischen Mittels berechnete mittlere Bevölkerungszahl genau übereinstimmen mit der in der Mitte des ganzen Zeitraums durch Zählung festgestellten Personenzahl

Gerade umgekehrt liegen die Verhältnisse im Zeitraum 1925—1933. Bei der Volkszählung am 16. Juni 1925 wurden im Deutschen Reich 62 410 619 Personen gezählt, bei der am 16. Juni 1933 65 218 461 [2]). Das arithmetische Mittel berechnet sich auf 63 814 540, das geometrische auf 63 799 100. Schreibt man auf Grund der Statistik der Geburten, Sterbefälle und Wanderungen die Volkszählungszahl von 1925 fort, so erhält man für Mitte 1929 eine Bevölkerungszahl von 63 968 970 [3]). Das arithmetische Mittel liegt etwas näher an dieser fortgeschriebenen Zahl als das geometrische Mittel. Somit ist in diesem Zeitraum, in dem die Reichsbevölkerung nur langsam wuchs, die Anwendung des arithmetischen Mittels zweckmäßiger als die des geometrischen Mittels.

[1]) Statistisches Jahrbuch für das Deutsche Reich 1900, S. 1; 1903, S. 1.

[2]) Statistisches Jahrbuch für das Deutsche Reich 1936, S. 5.

[3]) Statistisches Jahrbuch für das Deutsche Reich 1936, S. 5 u. 7.

Allgemein läßt sich sagen, daß die Berechnung der mittleren Bevölkerung in Zeiten raschen Bevölkerungswachstums sachgemäß unter Anwendung des geometrischen Mittels und in Zeiten langsamen Wachstums unter Anwendung des arithmetischen Mittels erfolgt. Dies gilt im besonderen für die Berechnung der mittleren Bevölkerungszahl für ein Kalenderjahr. Man bestimmt durch Fortschreibung die Bevölkerungszahl am Anfang und am Ende des Kalenderjahres und zieht aus beiden Bestandszahlen das Mittel. In der amtlichen Statistik wurde der vorstehenden Erwägung zufolge die Berechnung der mittleren Bevölkerungszahl eines Kalenderjahres vor dem Kriege mit dem geometrischen Mittel bewirkt, während diese Berechnung gegenwärtig mit dem arithmetischen Mittel durchgeführt wird.

Hierzu sei noch das Folgende bemerkt. Die exakteste Zahl für die mittlere Bevölkerung in einem Kalenderjahr ist zweifelsohne die Zahl der Jahre, die von der Gesamtheit der Personen im Kalenderjahr verlebt worden sind. Bezeichnen wir entsprechend den oben eingeführten Abkürzungen die Bevölkerungszahl am Anfang des Kalenderjahres mit V_0 und am Ende des Kalenderjahres mit V_1 und zu einem beliebigen Zeitpunkt t während des Kalenderjahres mit V_t, wobei $0 \leqq t \leqq 1$ ist, so ergibt sich für die verlebte Zeit die Integraldarstellung

$$\text{Verlebte Zeit} = \int_0^1 V_t \, dt. \tag{24}$$

Die Maßzahl der verlebten Zeit läßt sich ohne weiteres für die beiden Sonderfälle der konstanten absoluten und der konstanten relativen Bevölkerungszunahme angeben.

Bei konstanter absoluter Bevölkerungszunahme beträgt die Bevölkerungszahl V_t zur Zeit t

$$V_t = V_0 + (V_1 - V_0)\, t.$$

Führt man diesen Ausdruck für V_t in das Integral der verlebten Zeit (24) ein, so erhält man

$$\text{Verlebte Zeit} = \int_0^1 [V_0 + (V_1 - V_0)t]\, dt = \tfrac{1}{2}(V_0 + V_1).$$

Für die Mitte des Kalenderjahres ($t = \frac{1}{2}$) berechnet sich bei konstanter absoluter Bevölkerungszunahme die Bevölkerungszahl auf

$$V_{\frac{1}{2}} = \frac{V_0 + V_1}{2}.$$

Hieraus folgt, daß bei konstanter absoluter Bevölkerungszunahme die Maßzahl für die verlebte Zeit gleich ist der Bevölkerungszahl in der Mitte des Kalenderjahres und daß beide Zahlenwerte gleich dem arithmetischen Mittel aus der Anfangs- und Endbevölkerung sind.

Den zweiten Sonderfall der konstanten relativen Bevölkerungszunahme wollen wir durch die Konstanz der Vermehrungsintensität ρ_t kennzeichnen, wobei

$$\rho_t = \frac{1}{V_t} \frac{dV_t}{dt}$$

ist. Bei konstantem ρ_t ergibt sich

$$V_t = V_0 \left(\frac{V_1}{V_0}\right)^t.$$

Setzt man den Ausdruck für V_t in das Integral für die verlebte Zeit (24) ein, so erhält man

$$\text{Verlebte Zeit} = \frac{V_1 - V_0}{ln\, V_1 - ln\, V_0}.$$

Für die Mitte des Kalenderjahres ($t = \frac{1}{2}$) berechnet sich die Bevölkerungszahl $V_{\frac{1}{2}}$ bei konstanter relativer Bevölkerungszunahme auf

$$V_{\frac{1}{2}} = \sqrt{V_0\, V_1}.$$

Bei konstanter relativer Bevölkerungszunahme ist somit die verlebte Zeit gleich dem absoluten Bevölkerungszuwachs, dividiert durch den logarithmischen Bevölkerungszuwachs, und die Bevölkerungszahl in der Mitte des Kalenderjahres ist gleich dem geometrischen Mittel aus der Anfangs- und Endbevölkerung.

Bei konstanter relativer Bevölkerungszunahme ist allgemein das geometrische Mittel kleiner als die verlebte Zeit, da die Ungleichung gilt

$$\sqrt{V_0\, V_1} < \frac{V_1 - V_0}{ln\, V_1 - ln\, V_0} \tag{25}$$

Die Richtigkeit dieser Ungleichung kann in folgender Weise bewiesen werden. Man setzt

$$\frac{V_0}{V_1} = 1 + x. \tag{26a}$$

Hierbei ist $-1 < x < 0$, da $V_0 < V_1$ ist. Dividiert man beide Seiten der letzten Ungleichung durch V_1, so gelangt man zu der Form

$$\sqrt{1+x} < \frac{x}{ln\,(1+x)}.$$

Hieraus ergibt sich

$$ln\,(1+x) > \frac{x}{\sqrt{1+x}}.$$

Für $|x| < 1$ lassen sich die beiden folgenden Reihen ansetzen:

$$ln\,(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} \pm \cdots, \tag{27}$$

$$\frac{x}{\sqrt{1+x}} = x - \frac{1}{2}\, x^2 + \frac{1 \cdot 3}{2 \cdot 4}\, x^3 - \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6}\, x^4 \pm \cdots \tag{28}$$

Vergleicht man die Koeffizienten entsprechender Potenzen von x in den beiden Reihen (27) und (28), so findet man, da alle Reihenglieder negativ sind, die Richtigkeit der Ungleichung (25).

Die gleiche Betrachtung läßt sich auch für den Fall der konstanten relativen Bevölkerungsabnahme durchführen. In diesem Falle ist zu setzen:

$$\frac{V_1}{V_0} = 1 + x. \tag{26b}$$

Weiter kann man zeigen, daß bei konstanter relativer Bevölkerungszunahme das arithmetische Mittel größer ist als die verlebte Zeit. Zum Nachweis der entsprechenden Ungleichung

$$\frac{V_0 + V_1}{2} > \frac{V_1 - V_0}{ln\, V_1 - ln\, V_0} \tag{29}$$

verwenden wir die Substitution (26a). Mit ihrer Hilfe erhält man

$$1 + \frac{x}{2} > \frac{x}{ln(1+x)}.$$

Eine einfache Umformung führt zu

$$ln(1+x) < \frac{x}{1+\frac{x}{2}}.$$

Da $|x| < 1$ ist, läßt sich die Funktion auf der linken Seite der Ungleichung durch die bereits angegebene konvergente Reihe (27) und die Funktion auf der rechten Seite durch die folgende Reihe darstellen:

$$\frac{x}{1+\frac{x}{2}} = x - \frac{x^2}{2} + \frac{x^3}{4} - \frac{x^4}{8} \pm \tag{30}$$

Vergleicht man die Koeffizienten gleicher Potenzen von x in den beiden Reihen (27) und (30), so ergibt sich die Richtigkeit der Ungleichung (29). Diese Ungleichung gilt auch im Falle der konstanten relativen Bevölkerungsverminderung. Wir wenden dann die Substitution (26b) an.

Es ist noch die Frage zu diskutieren, ob für den Sonderfall der konstanten relativen Bevölkerungszunahme das geometrische Mittel oder das arithmetische Mittel näher an der verlebten Zeit liegt. Stellen wir zu diesem Ende die Abstände des arithmetischen und des geometrischen Mittels von der verlebten Zeit dar, so gelangen wir zu der Ungleichung

$$\frac{V_0 + V_1}{2} - \frac{V_1 - V_0}{ln\, V_1 - ln\, V_0} > \frac{V_1 - V_0}{ln\, V_1 - ln\, V_0} - \sqrt{V_0 V_1} \tag{31}$$

Zum Nachweis der Richtigkeit dieser Ungleichung setzen wir wiederum die Substitution (26a) an. Wir erhalten auf diese Weise

$$\left(1+\frac{x}{2}\right) ln(1+x) - x < x - \sqrt{1+x}\, ln(1+x).$$

Da $|x| < 1$ ist, können wir $\sqrt{1+x}$ in folgende Reihe entwickeln:

$$\sqrt{1+x} = 1 + \frac{1}{2}x - \frac{1}{2\cdot 4}x^2 + \frac{1}{2\cdot 4}\cdot\frac{3}{6}x^3 \mp \cdots$$

Stellen wir $ln(1+x)$ durch die Reihe (27) dar, so gewinnen wir die Ungleichung

$$\left(1+\frac{x}{2}\right)\left(x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}\pm\cdots\right)-x<$$

$$<x-\left(1+\frac{1}{2}x-\frac{1}{2\cdot 4}x^2+\frac{1}{2\cdot 4\cdot 6}x^3\mp\cdots\right)\left(x-\frac{x^2}{2}+\frac{x^3}{3}\mp\cdots\right).$$

Durch die Koeffizientenvergleichung überzeugen wir uns sofort von der Richtigkeit der Ungleichung (31).

Dieselbe Ungleichung gilt auch für den Fall der konstanten relativen Bevölkerungsabnahme, was unter Anwendung der Substitution (26b) gezeigt werden kann.

Zusammenfassend stellen wir fest, daß bei konstanter absoluter Bevölkerungszunahme bzw. -abnahme das arithmetische und das geometrische Mittel gleich der Maßzahl für die verlebte Zeit sind und daß bei konstanter relativer Bevölkerungszunahme das geometrische Mittel näher an der Maßzahl für die verlebte Zeit liegt als das arithmetische Mittel. Nach bevölkerungsstatistischen Beobachtungen liegt in Zeiten raschen Bevölkerungswachstums nahezu der Fall der konstanten relativen Bevölkerungszunahme und in Zeiten langsamen Bevölkerungswachstums der Fall der konstanten absoluten Bevölkerungszunahme vor. Somit gelangen wir auf mathematischem Wege zu der oben empirisch festgestellten Tatsache, daß in Zeiten raschen Bevölkerungswachstums, wie wir sie vor dem Kriege hatten, zur Bestimmung der mittleren Bevölkerung das geometrische Mittel geeigneter ist als das arithmetische Mittel und in Zeiten langsamen Bevölkerungswachstums, wie wir sie gegenwärtig haben, das arithmetische Mittel zweckmäßiger ist als das geometrische [1]).

48. Mit dem geometrischen Mittel hängt eine Behandlungsweise der Kollektive zusammen, die Fechner unter dem Namen der logarithmischen neben die bisher besprochene arithmetische gestellt hat. Sie besteht, um es kurz zu sagen, in der Betrachtung von Verhältnissen statt von Differenzen. Eine Begründung für die Wahl dieses Vorgangs ist schon im vorigen Artikel gegeben worden; eine andere besteht darin, daß es Verteilungen gibt, die sich der logarithmischen Behandlung besser anpassen als der arithmetischen; das ist z. B. der Fall bei sehr großer Ausbreitung eines Kollektivs im Vergleich zu dem Mittelwert, auf den man sich beziehen will. Indessen braucht die Anwendung der Methode nicht auf solche Fälle allein beschränkt zu werden.

Das logarithmische Verfahren besteht nun im wesentlichen darin, daß man nicht die Argumentwerte selbst, sondern deren Logarithmen zur Grundlage der Klassenbildung macht. Bildet man auch jetzt gleiche Klassenintervalle, so entspricht dieser gleichmäßigen logarithmischen Skala eine ungleichmäßige der Argumente und somit eine andere Häufigkeitsverteilung, was unter Umständen Vorteile bieten kann.

Man kann auf die so hergestellte logarithmische Verteilungstafel alle Prozesse und genau so anwenden, wie sie an der arithmetischen Verteilungstafel geübt

[1]) Vgl. F. Burkhardt, Die Bewegung der Sterblichkeit im Spiegel der Statistik, Assekuranz-Jahrbuch 1933, S. 192 u. f.

worden sind, und gelangt so zum Teil zu den bereits bekannten, zum Teil zu neuen Größen, wie die folgenden Überlegungen zeigen werden.

1. Das arithmetische Mittel $\mathfrak{M}$ ist der Logarithmus des geometrischen Mittels G; somit erhält man dieses, indem man von $\mathfrak{M}$ zum Numerus übergeht.

2. Der Zentralwert $\mathfrak{C}$ ist der Logarithmus des früheren Zentralwertes C; denn $\mathfrak{C}$ scheidet die logarithmischen Argumentwerte in zwei gleich besetzte Gruppen, C bewirkt dieselbe Scheidung bei den Argumentwerten selbst; somit entspricht die Stellung des $\mathfrak{C}$ in der logarithmischen Reihe genau der Stellung von C in der arithmetischen Reihe.

3. Der dichteste Wert $\mathfrak{D}$ bezeichnet die Stelle, um welche sich die logarithmischen Argumentwerte häufen; den Differenzen in Bezug auf $\mathfrak{D}$ entsprechen Verhältniszahlen in Bezug auf den Numerus zu $\mathfrak{D}$, der mit D_v bezeichnet werden soll; mithin ist D_v im arithmetischen Gebiet diejenige Stelle, um welche sich die *Verhältnisse* der Argumentwerte zum Vergleichswert (d. i. zu G) am dichtesten zusammendrängen, und soll darum der dichteste Verhältniswert heißen.

Die folgenden Beispiele sollen die logarithmische Behandlung von Kollektiven verständlich machen und ihr Verhältnis zur arithmetischen ins rechte Licht stellen.

Beispiel 1). Das Vorzuführende ist eine Neubearbeitung der Regenhöhen der einzelnen Regentage des Monats Januar in Genf während der 48 Jahre von 1845 bis 1892[1]). Die arithmetische Verteilung gestaltet sich so:

Tab. 38. Regenhöhen im Januar 1845 bis 1892 in Genf.

X in mm	z Zahl der Regentage	X in mm	z Zahl der Regentage	X in mm	z Zahl der Regentage
0— 1	133	14—15	3	28—29	1
1— 2	88	15—16	3	29—30	.
2— 3	43,5	16—17	2	30—31	1
3— 4	28	17—18	5	31—32	.
4— 5	27	18—19	1	32—33	1
5— 6	28	19—20	3	33—34	.
6— 7	27,5	20—21	.	34—35	.
7— 8	14,5	21—22	3	35—36	.
8— 9	16	22—23	.	36—37	.
9—10	11,5	23—24	2	37—38	.
10—11	12	24—25	.	38—39	.
11—12	10	25—26	.	39—40	1
12—13	6,5	26—27	.		
13—14	5,5	27—28	.		477

[1]) G. Th. Fechner, Kollektivmaßlehre, Leipzig 1897, S. 344 u. f.

Die Tafel führt zu folgenden Mittelwerten:

$$M = 4{,}486, \quad C = 2{,}403, \quad D = 0;$$

solange eine Aufteilung der ersten Klasse fehlt, erscheint 0 als Häufungsstelle, also als dichtester Wert, die Verteilung ist eine einseitige. Ihre Ausbreitung ist ein großes Vielfaches, fast das Neunfache des arithmetischen Mittels; gegen das Ende zu werden leere Klassen immer zahlreicher, eine größere Zusammenfassung des oberen Teils, dafür eine feinere Zergliederung des untern ist erwünscht; beides wird durch die logarithmische Behandlung erzielt, die an die Urliste anknüpft.

Tab. 38 in logarithmischer Behandlung.

log X	X in mm	Arithm. Klassengröße	z	v	Differenz
— 1,5 bis — 1,3	0,03— 0,05	0,02	8	0,006	0,004
— 1,3 „ — 1,1	0,05— 0,08	0,03	8	0,010	0,006
— 1,1 „ — 0,9	0,08— 0,13	0,05	9	0,016	0,009
— 0,9 „ — 0,7	0,13— 0,20	0,07	9	0,025	0,015
— 0,7 „ — 0,5	0,20— 0,32	0,12	28	0,040	0,023
— 0,5 „ — 0,3	0,32— 0,50	0,18	14	0,063	0,037
— 0,3 „ — 0,1	0,50— 0,79	0,29	34	0,100	0,04
— 0,1 „ 0,1	0,79— 1,26	0,47	45	0,16	0,09
0,1 „ 0,3	1,26— 2,00	0,74	66	0,25	0,15
0,3 „ 0,5	2,00— 3,16	1,16	47	0,40	0,23
0,5 „ 0,7	3,16— 5,01	1,85	53	0,63	0,37
0,7 „ 0,9	5,01— 7,94	2,93	67	1,00	0,58
0,9 „ 1,1	7,94—12,59	4,65	53	1,58	0,93
1,1 „ 1,3	12,59—19,95	7,36	27	2,51	1,47
1,3 „ 1,5	19,95—31,62	11,67	7	3,98	2,33
1,5 „ 1,7	31,62—50,12	18,50	2	6,31	
			477		

Das konstante logarithmische Intervall ist mit 0,2 festgesetzt worden, neben die logarithmischen Intervalle sind die arithmetischen geschrieben, um zu zeigen, daß den gleichen logarithmischen Klassen wachsende arithmetische Klassen entsprechen. Die arithmetische Klassengröße beginnt jetzt mit 0,02 und steigt bis auf 18,50 mm. Ordnet man einer Klasse jene Verhältniszahl zu, welche zur Mitte der logarithmischen Klasse gehört, so ergeben sich die Zahlen der Kolonne v, und die letzte mit Differenz überschriebene Kolonne zeigt, wie diese Verhältniszahlen von Klasse zu Klasse anwachsen. Die Berechnung der Kolonne v erfolgt hiernach so, daß man die Differenzen

$$-1{,}4 - \mathfrak{D}, \qquad -1{,}2 - \mathfrak{D}, \qquad -1{,}0 - \mathfrak{D},$$

bildet und dazu die Numeri nimmt; die Werte in den ersten 11 Klassen sind kleiner als 1, der Wert der 12. Klasse, in der $\mathfrak{D}$ liegt, ist gleich 1, und die Werte in den weiteren Klassen sind größer als 1.

Die drei logarithmischen Mittelwerte sind:

$$\mathfrak{M} = 0{,}314, \quad \mathfrak{C} = 0{,}374, \quad \mathfrak{D} = 0{,}8;$$

dazu gehören die Numeri: $G = 2{,}061, \quad C = 2{,}366, \quad D_c = 6{,}310$ mm.

Das Bild ist ein völlig verändertes; das Kollektiv ist von 40 auf 16 Klassen zusammengezogen; an die Stelle der einseitigen Verteilung ist eine zweiseitige mit starker rechtsseitiger Asymmetrie getreten; die Anfangsklassen sind feiner aufgeteilt, die Endklassen summarisch zusammengefaßt. Die beobachtete Erscheinung unterliegt einer starken Variation, denn das Verhältnis $\frac{X}{D_v}$ durchläuft das Intervall 0,006 bis 6,31.

Beispiel 2). In der Tab. 39 werden die Monatssummen des Niederschlages in Dresden im Januar in den Jahren 1828 bis 1933[1]) in Gruppen aufgeteilt mit Angabe der Zahl der Jahre, in denen die Niederschlagsmenge in der betreffenden Gruppe lag.

Die drei Mittelwerte stellen sich auf

$$M = 35{,}5;$$
$$C = 32{,}7;$$
$$D = 29{,}0.$$

Die Verteilung ist somit linksseitig asymmetrischer Natur.

Geht man von der arithmetischen zur logarithmischen Behandlung über, so erhält man das folgende Zahlenbild:

Tab. 39. Monatssumme des Niederschlages in Dresden im Januar 1828—1933.

X in mm	z Zahl der Jahre
4,0— 14,0	14
14,0— 24,0	19
24,0— 34,0	23
34,0— 44,0	19
44,0— 54,0	13
54,0— 64,0	10
64,0— 74,0	3
74,0— 84,0	3
84,0— 94,0	0
94,0—104,0	2
	106

Tab. 39 in logarithmischer Behandlung.

log X	X in mm	Arithm. Klassengröße	z	v	Differenz
0,6—0,8	3,98— 6,31	2,33	2	0,13	
					0,08
0,8—1,0	6,31— 10,00	3,69	7	0,21	
					0,12
1,0—1,2	10,00— 15,85	5,85	10	0,33	
					0,19
1,2—1,4	15,85— 25,12	9,27	17	0,52	
					0,30
1,4—1,6	25,12— 39,81	14,69	31	0,82	
					0,48
1,6—1,8	39,81— 63,10	23,29	30	1,30	
					0,76
1,8—2,0	63,10—100,00	36,90	9	2,06	
			106		

[1]) Festschrift der Landeswetterwarte, 16. Tagung zu Dresden. Dresden 1929, S. XVIII; Deutsches Meteorologisches Jahrbuch 1928—1933.

In dieser Tabelle ist die logarithmische Klassengröße mit 0,2 angenommen worden.

Die drei logarithmischen Mittelwerte sind:

$$\mathfrak{M} = 1{,}466, \qquad \mathfrak{C} = 1{,}510, \qquad \mathfrak{D} = 1.587.$$

Dazu gehören die Numeri:

$$G = 29{,}242, \qquad C = 32{,}359, \qquad D_v = 38{,}637.$$

Bei arithmetischer Bearbeitung ist $C = 32{,}7$.

Das geometrische Mittel liegt beträchtlich unter dem arithmetischen, während der dichteste Verhältniswert D_v den dichtesten Wert D übertrifft. Beim Übergang von der arithmetischen zur logarithmischen Darstellung wandelt sich die linksseitige Asymmetrie in eine rechtsseitige Asymmetrie um.

49. Das harmonische Mittel. Hat man den Häufigkeitszahlen z nicht die Argumentwerte X selbst, sondern deren Reziproke $\frac{1}{X}$ zugeordnet, so heißt das Reziproke des auf dieser Grundlage gebildeten arithmetischen Mittels das harmonische Mittel der X; gibt man ihm das Zeichen H, so besteht dafür die Definition:

$$\frac{1}{H} = \frac{1}{N} \Sigma \left(\frac{z}{X} \right). \tag{32}$$

Beispiel. Auf 1000 Einwohner kamen am 1. Juli 1935 Kraftfahrzeuge[1]) in:

1.	Ostpreußen	23
2.	Stadt Berlin	37
3.	Provinz Brandenburg	40
4.	„ Pommern	29
5.	„ Grenzmark Posen-Westpreußen	29
6.	„ Niederschlesien	32
7.	„ Oberschlesien	16
8.	„ Sachsen	37
9.	„ Schleswig-Holstein	34
10.	„ Hannover	33
11.	„ Westfalen	24
12.	„ Hessen-Nassau	32
13.	Rheinprovinz	27

Man kann daraus auf zwei Arten berechnen, wieviel Personen im Mittel auf ein Kraftfahrzeug kommen; einmal, indem man bestimmt, wieviel Kraftfahrzeuge im Mittel in einer Provinz auf 1000 Einwohner entfallen:

[1]) Vierteljahrshefte zur Statistik des Deutschen Reichs 1935, III, S. 52.

$$\frac{393}{13} = 30{,}23$$

und daraus die Zahl der Personen pro Kraftfahrzeug

$$\frac{1000}{30{,}23} = 33{,}08\,;$$

oder indem man ansetzt

$$\frac{1}{H} = \frac{1}{13}\left(\frac{1}{16} + \frac{1}{23} + \frac{1}{24} + \frac{1}{27} + \frac{2}{29} + \frac{2}{32} + \frac{1}{33} + \frac{1}{34} + \frac{2}{37} + \frac{1}{40}\right) = 0{,}03492,$$

woraus sich die Zahl der Personen pro Kraftfahrzeug zu 34,92, also höher als vorhin ergibt. Das aus dem letzten Ansatz resultierende H ist das harmonische Mittel der Anzahl der Kraftfahrzeuge auf 1000 Einwohner. Es liegt mit 28,64 etwas niedriger als das arithmetische[1]). Bei der ersten Rechnung ist aus dem arithmetischen Mittel der Zahl der Kraftfahrzeuge auf 1000 Einwohner auf die Zahl der Personen pro Kraftfahrzeug, bei der zweiten aus dem arithmetischen Mittel der Personen pro Kraftfahrzeug auf die Zahl der Kraftfahrzeuge auf 1000 Einwohner geschlossen worden.

Im Vorstehenden ist mit dem einfachen arithmetischen und dem einfachen harmonischen Mittel gearbeitet worden. Weichen die Einwohnerzahlen der verschiedenen räumlichen Gebiete stark voneinander ab, so ist es notwendig, das gewogene arithmetische und das gewogene harmonische Mittel zu bilden. Hierbei führt man zweckmäßig die statistischen Gewichte in der Weise ein, daß man das kleinste räumliche Gebiet als Einheit ansetzt und die übrigen Gebiete nach Maßgabe dieses Einheitsgebietes in Teilgebiete zerlegt. Zahlenmäßig geschieht dies so, daß man das Verhältnis der Einwohnerzahlen der größeren Gebiete zur Einwohnerzahl des Einheitsgebietes bestimmt und die so erhaltenen Verhältniszahlen als z-Werte in die Rechnung einführt (vgl. hierzu Art. 87).

50. Das quadratische Mittel. Im Hinblick auf die in § 4 dieses Abschnittes anzustellenden Betrachtungen über die Streuungsmaße sei hier kurz noch das quadratische Mittel eingeführt. Sind $X_1, X_2, \ldots X_n$ die einzelnen Werte des Merkmals X, so berechnet man das quadratische Mittel in folgender Weise:

$$\text{Quadratisches Mittel} = \sqrt{\frac{1}{n}\left(X_1^2 + X_2^2 + \ldots X_n^2\right)} = \sqrt{\frac{1}{n}\Sigma X_i^2}$$

Das quadratische Mittel tritt auch in der Mechanik auf, u. zw. als Trägheitsradius. Bei der Berechnung des Trägheitsradius geht man von den Abständen der einzelnen Massenpunkte eines Systems von einer gegebenen Achse aus und bestimmt das quadratische Mittel dieser Abstände. In der Statistik spielt das quadratische Mittel, wie schon erwähnt, in der Dispersionslehre eine wichtige Rolle.

[1]) Das arithmetische Mittel zweier positiven Größen a, b $(a \neq b)$ ist $M = \frac{a+b}{2}$, das harmonische Mittel $H = \frac{2ab}{a+b}$; $M - H = \frac{(a-b)^2}{2(a+b)}$, daher $M > H$.

§ 3. Verhältniszahlen.

51. Begriff und Arten der Verhältniszahlen. Bei der Bildung von Verhältniszahlen werden zwei verschiedene statistische Massen zueinander in Beziehung gesetzt. Wir wollen die Bildung von Verhältniszahlen an dem Beispiel der Industriequote erläutern. Bei ihrer Berechnung geht man von der Zahl der Erwerbspersonen (Erwerbstätige und Arbeitslose) aus, stellt fest, wie viele von ihnen der Industrie (einschließlich Handwerk) zuzuordnen sind und berechnet den relativen Anteil der letzteren an der Gesamtzahl der Erwerbspersonen. Bezeichnet man die Zahl der Erwerbspersonen mit e und die Zahl der auf die Industrie entfallenden mit i, so ist die

$$\text{Industriequote} = \frac{i}{e}.$$

Nach der Volks- und Berufszählung im Deutschen Reich am 16. Juni 1933 [1]) betrug die Zahl der Erwerbspersonen 32 296 074 und die Zahl der zur Industrie gehörigen 13 052 982. Danach ergibt sich eine Industriequote von 0,40.

In der praktischen Statistik werden vielfach die Verhältniszahlen, um sie leicht lesbar, schreibbar und vorstellbar zu machen, mit 100 oder 1000 oder auch mit einer höheren Potenz von 10 multipliziert. Dadurch wird bewirkt, daß die bei der Quotenbildung im Zähler stehende Zahl (Zählermasse) nicht auf die Einheit, sondern auf 100, 1000 ... Einheiten der Nennermasse bezogen wird. Die mit 100 bzw. 1000 multiplizierten Verhältniszahlen erhalten den Zusatz „v. H.“ bzw. „v. T.“. Die Industriequote des Deutschen Reiches schreibt man z. B. vielfach in der Form 40 v. H.

Im Hinblick auf das Verhältnis der Zähler- und Nennermassen zueinander lassen sich, der Terminologie von G. v. Mayr [2]) und Žižek [3]) im allgemeinen folgend, drei Arten von Verhältniszahlen unterscheiden: Gliederungszahlen, Beziehungszahlen und Meßzahlen.

52. Gliederungszahlen. Von Gliederungszahlen spricht man, wenn die Zählermasse einen Teil der Nennermasse bildet. Dies ist z. B. bei der bereits behandelten Industriequote der Fall. Nach Lexis [4]) und v. Bortkiewicz [5]) bezeichnet man die Gliederungszahlen auch als analytische Verhältniszahlen, weil man bei der Bildung von Gliederungszahlen die Grund- oder Nennermasse zerlegt und die durch Zerlegung erhaltenen Teilgesamtheiten zur Grundmasse ins Verhältnis setzt.

Die Methode der Gliederungszahlen verwendet man in der praktischen Statistik, um die Struktur von Kollektiven zu kennzeichnen. Die bereits angeführte Industriequote läßt zusammen mit der Landwirtschafts- und der Handelsquote (pro-

[1]) Statistisches Jahrbuch für das Deutsche Reich 1936, S. 18 u. 19.

[2]) G. v. Mayr, Theoretische Statistik, Tübingen 1914, S. 156 u. f.

[3]) F. Žižek, Grundriß der Statistik, München und Leipzig 1923, S. 134 u. f.

[4]) W. Lexis, Abhandlungen zur Theorie der Bevölkerungs- und Moralstatistik. Jena 1903, S. 84.

[5]) L. v. Bortkiewicz, Grundriß einer Vorlesung über Allgemeine Theorie der Statistik. Berlin 1907, S. 16.

zentualer Anteil der in der Landwirtschaft bzw. im Handel Gezählten an der Gezamtzahl der Erwerbspersonen) die wirtschaftliche Struktur eines Landes erkennen. Auf Grund dieser Quoten wird es möglich, Industrie- und Agrarländer zu unterscheiden. Ebenso wird für das Personenkollektiv der Volkszählung die Alters-, Familienstands-, Geschlechts-, Religions- usw. -gliederung durch analytische Verhältniszahlen dargestellt. Die Leichtverständlichkeit und Anschaulichkeit der Gliederungszahlen, namentlich in der Schreibweise v. H., v. T., hat sie zu einem der verbreitetsten statistischen Ausdrucksmittel gemacht. Dies gilt nicht nur von den konkreten Kollektiven, sondern auch von den abstrakten konstruierten statistisch-volkswirtschaftlichen Zahlenwerten, so z. B. vom Volksvermögen und Volkseinkommen. Nach F. Zahn[1]) besteht das gesamte Volksvermögen mindestens zu 75 v. H. aus dem organischen Volkskapital, und das Volkseinkommen kann zu 80 bis 90 v. H. als unmittelbarer Ertrag der menschlichen Arbeit gelten.

53. Beziehungszahlen. Ist die Zählermasse nicht ein Teil der Nennermasse, sondern stehen sich die beiden Massen koordiniert gegenüber, so bezeichnet man die durch Inbeziehungsetzen beider entstehenden Verhältniszahlen als Beziehungszahlen. Als Beispiel sei die Eheschließungsziffer angeführt. Zu ihrer Bestimmung setzt man die Zahl der Eheschließungen in einem bestimmten Zeitraum und in einem bestimmten örtlichen Gebiet zur Gesamtbevölkerung in Beziehung. Es sind zwei Arten von Beziehungszahlen zu unterscheiden: genetische Beziehungszahlen und Entsprechungszahlen.

a) Den genetischen Beziehungszahlen liegt ein reales Geschehen zugrunde, das sich an den Elementen der Nennermasse vollzieht. Die Ereignisfälle dieses Geschehens bilden die Zählermasse. Die Elemente derselben werden gleichsam von den Elementen der Nennermasse erzeugt. Zu den genetischen Beziehungszahlen gehören z. B. die Eheschließungs-, Geburten- und Sterbeziffern. Je nach der Methode, nach der die Berechnung erfolgt, unterscheidet man die statistische Wahrscheinlichkeit und den statistischen Koeffizienten.

Statistische Wahrscheinlichkeiten erhält man, wenn man die Zahl der Ereignisfälle innerhalb eines Zeitraumes zum Bestand der Nennermasse am Anfang des Zeitraums in Beziehung setzt, und statistische Koeffizienten gewinnt man, wenn man die Zahl der Ereignisfälle innerhalb eines Zeitraums zum mittleren Bestand der Nennermasse ins Verhältnis bringt. Bezeichnet man die Bestandszahl der Nennermasse zu Beginn des Zeitraums x bis $x+1$ mit r_x und die mittlere Bestandszahl mit m_x, sowie die Zahl der Ereignisfälle im Zeitraum x bis $x+1$ mit s_x, so stellt sich die statistische Wahrscheinlichkeit q_x in der Form

$$q_x = \frac{s_x}{r_x} \tag{1a}$$

und der statistische Koeffizient k_x in der Form

$$k_x = \frac{s_x}{m_x} \tag{1b}$$

[1]) F. Zahn, Das Bevölkerungsproblem und die volkswirtschaftliche Kapitalbildung. Bevölkerungsfragen, Bericht des Internationalen Kongresses für Bevölkerungswissenschaft in Berlin 1935. München 1936, S. 231 u. f.

dar. Zur Bestimmung der mittleren Bestandszahl bezeichnen wir die Bestandszahl zu einem beliebigen Zeitpunkt $x+t$ $(0 \leqq t \leqq 1)$ mit r_{x+t}. Auf diese Weise erhalten wir für die mittlere Bestandszahl die Integraldarstellung

$$m_x = \int^{1} r_{x+t}\, dt, \tag{2}$$

und somit ergibt sich für den statistischen Koeffizienten k_x die Form

$$k_x = \frac{s_x}{\int^{1} r_{x+t}\, dt}. \tag{3}$$

Als Beispiel hierzu sei kurz die Sterbenswahrscheinlichkeit und der Sterblichkeitskoeffizient behandelt. Ist r_x die Zahl der Lebenden im Alter x und r_{x+1} die entsprechende Zahl für das Alter $x+1$, sowie s_x die Zahl der Gestorbenen im Alter x bis $x+1$, so ist die Sterbenswahrscheinlichkeit durch (1a) und der Sterblichkeitskoeffizient durch (3) gegeben.

Verändert sich der Bestand der Nennermasse im Laufe der Zeit absolut gleichmäßig, d. h. ist r_{x+t} eine lineare Funktion von t, z. B.

$$r_{x+t} = r_x - t(r_x - r_{x+1}),$$

so ist

$$\int_0^1 r_{x+t}\, dt = r_x - \frac{r_x - r_{x+1}}{2}$$

$$k_x = \frac{s_x}{r_x - \frac{r_x - r_{x+1}}{2}}. \tag{3a}$$

Für den Sterblichkeitskoeffizienten erhält man unter der Annahme, daß die Zahl der Lebenden r_x durch Abzug der Gestorbenen s_x in die Lebendenzahl r_{x+1} übergeht, die folgende Form

$$k_x = \frac{q_x}{1 - \frac{q_x}{2}}. \tag{3b}$$

b) Die Entsprechungszahlen dagegen drücken eine willkürlich und künstlich hergestellte Beziehung zwischen begrifflich nicht verwandten Massen aus. Bei der Zuordnung, deren Sinn die Ermittlung eines gewissen Verteilungsdurchschnitts ist, können Zähler- und Nennermasse willkürlich vertauscht werden, d. h. man kann sowohl von der einen als auch von der anderen Masse ausgehen, wie z. B. bei der allgemeinen Bevölkerungsdichteziffer, die durch Inbeziehungsetzen der Bevölkerungszahl zur Gebietsfläche gewonnen wird und besagt, wieviel Einwohner durchschnittlich auf 1 km² wohnen. Ebenso kann man die Gebietsfläche zur Einwohnerzahl in Beziehung setzen und ermitteln, wieviel Quadratkilometer im Durchschnitt einem Einwohner zur Verfügung stehen. Die Umkehrbarkeit der Zuordnung liegt in der Wesensfremdheit und gegenseitigen Unabhängigkeit der beiden Massen begründet.

Im Hinblick darauf, daß in verschiedenen Ländern nicht die gesamte Gebietsfläche des Landes bewohnbar ist, hat man die Frage aufgeworfen, ob es nicht richtiger wäre, die Bevölkerungszahl auf die bewohnbare Fläche (Siedlungsfläche) zu beziehen. Mit diesem Problem hat sich J. Müller [1]) eingehend beschäftigt. Müller legt dar, daß der Begriff Siedlungsfläche nicht eindeutig ist. Man kann ihn eng und weit umgrenzen. Bei enger Umgrenzung wird man Seen, Hochgebirge, Sümpfe und gegebenenfalls Waldland von der Gesamtfläche absetzen, bei weiter Umgrenzung wird man nur die Seenflächen in Abzug bringen. Zur Lösung dieser Schwierigkeiten schlägt Müller vor, für Vergleiche größerer Länder und Landesteile untereinander, namentlich in geographischer und bevölkerungswissenschaftlicher Hinsicht, die allgemeine Bevölkerungsdichteziffer zu verwenden und für wirtschaftswissenschaftliche Sonderuntersuchungen besondere Dichteziffern zu berechnen. Bei den besonderen Dichteziffern setzt man die Gesamtbevölkerung zur Siedlungsfläche in Beziehung. Hierbei ist darauf zu achten, daß die angesetzten Siedlungsflächen nach „Ausschaltung störender Teilmassen" vergleichbar sind.

Die Entsprechungszahlen spielen auch in der Verkehrsstatistik eine Rolle. Im städtischen Nahverkehr kennzeichnet O. G. A. Büchner [2]) das Verkehrsangebot durch die Zahl der gefahrenen Wagenkilometer und Platzkilometer und die Verkehrsnachfrage durch die Zahl der gefahrenen Personenkilometer. Diese drei zusammengesetzten statistischen Maßzahlen bezieht Büchner auf 1 km Streckenlänge. Durch Division der Zahl der gefahrenen Personenkilometer durch die Gesamtzahl der beförderten Personen ergibt sich die mittlere Reiselänge.

54. Meßzahlen. Unter Meßzahlen versteht man in der statistischen Methodenlehre solche Verhältniszahlen, die entstehen, wenn man ein Glied (Basisglied) einer statistischen Reihe gleich 1 oder gleich 100 oder gleich einer anderen Potenz von 10 setzt und die übrigen Reihenglieder dementsprechend umrechnet. Als Basisglied für diese Umrechnung wählt man entweder das Anfangs- oder das Endglied oder auch ein beliebiges Glied der Reihe. Zuweilen wird die Umrechnung auch in der Weise vorgenommen, daß man den Durchschnittswert der ganzen Reihe gleich 100 setzt. Das Meßzahlenverfahren wendet man auf den Gebieten der statistischen Forschung an, auf denen die Vergleichung statistischer Reihen von besonderer Wichtigkeit ist. Dies gilt vornehmlich von dem Gebiet der Konjunkturforschung. Hier verwendet man vielfach die konjunkturstatistischen Zahlen derjenigen Jahre als Basis, in denen die Konjunkturkurve ein Maximum oder ein Minimum durchlief.

Das Meßzahlenverfahren wendet man nicht bloß auf einfache, sondern auch auf zusammengesetzte statistische Zahlenreihen an. Von zusammengesetzten statistischen Zahlenreihen spricht man dann, wenn sich die einzelnen Glieder der Zahlenreihe auf mehrere statistische Größen beziehen, die in geeigneter Weise zusammengefaßt werden. Wie die Zusammenfassung erfolgen kann, wollen wir an dem Beispiel des Lebenshaltungsindex betrachten. Beim Lebenshaltungsindex geht man von einer fünfköpfigen Arbeiterfamilie aus, die aus dem Elternpaar und drei Kindern besteht. Man berechnet zunächst die Lebenshaltungskosten einer solchen

[1]) J. Müller, Wie wird die Bevölkerungsdichte richtig berechnet? Deutsches Statistisches Zentralblatt 1935, Heft 2, Sp. 33 u. f.

[2]) O. G. A. Büchner, Zur Methode der Feststellung von Angebot und Nachfrage im städtischen Nahverkehr. Bulletin de l'Institut international de Statistique. Tome XXVII, livraison 2, p. 482 u. f.

Familie in vier Wochen. Bei dieser Kostenberechnung werden die Ausgaben für Ernährung, Wohnung, Heizung, Beleuchtung, Bekleidung, Verkehr und für sonstigen Bedarf in Ansatz gebracht[1]). Im einzelnen führt man die Berechnung in der Weise durch, daß man für jedes Lebenshaltungsgut die Menge festlegt, die von der fünfköpfigen Familie in vier Wochen benötigt wird. So werde z. B. vom 1. Lebenshaltungsgut die Menge g_1, vom 2. die Menge g_2 u. s. f., vom n-ten die Menge g_n gebraucht. Die Zahlen $g_1, g_2, \ldots g_n$ tragen den Charakter von statistischen Gewichten. Sie werden durch die ganze Rechnung hindurch konstant gehalten. Die Preise der einzelnen Lebenshaltungsgüter seien der Reihe nach zu einem bestimmten Zeitpunkt T mit $p_1, p_2, \ldots p_n$ bezeichnet. Wir erhalten somit für die Lebenshaltungskosten zu einem bestimmten Zeitpunkt T den Ausdruck:

$$g_1 p_1 + g_2 p_2 + \cdots + g_n p_n = \Sigma g p.$$

Bestimmt man für verschiedene Zeitpunkte die Preise der einzelnen Lebenshaltungsgüter und mittels dieser Preise die Lebenshaltungskostenbeträge, so gelangt man zu einer zeitlichen statistischen Reihe von Lebenshaltungskostenzahlen. In dieser Reihe wählt man ein Glied als Basisglied. Die Zeit, auf die sich das Basisglied bezieht, bezeichnet man als Basiszeit. Die Preise zur Basiszeit seien $p'_1, p'_2, \ldots p'_n$. Das Basisglied stellt sich somit auf

$$g_1 p'_1 + g_2 p'_2 + \cdots + g_n p'_n = \Sigma g p'.$$

Bezieht man den Lebenshaltungskostenbetrag für den Zeitpunkt T (Messungszeit) auf das Basisglied, das gleich 100 gesetzt wird, so erhält man den Lebenshaltungsindex I zur Zeit T in der Form

$$I = 100 \frac{g_1 p_1 + g_2 p_2 + \cdots + g_n p_n}{g_1 p'_1 + g_2 p'_2 + \cdots + g_n p'_n} = 100 \frac{\Sigma g p}{\Sigma g p'}.$$

In der amtlichen Statistik wird als Basiszeit die Zeit 1913/14 angesetzt. Die Veröffentlichung[2]) erfolgt regelmäßig im Statistischen Jahrbuch für das Deutsche Reich und in Wirtschaft und Statistik.

Bei der Berechnung von Lebenshaltungsindizes treten dann Schwierigkeiten auf, wenn zur Messungszeit einzelne Lebenshaltungsgüter nicht mehr in denselben Mengen für die Lebenshaltung verwendet werden, wie zur Basiszeit und wenn einzelne Lebenshaltungsgüter, die zur Basiszeit noch wichtig waren, zur Messungszeit überhaupt nicht mehr hergestellt werden. Zur Behebung dieser beiden Schwierigkeiten bieten sich zwei Möglichkeiten. Die erste Möglichkeit besteht darin, daß man folgende Doppelformel für die Bestimmung des Lebenshaltungsindex ansetzt:

$$I = 100 \sqrt{\frac{\Sigma g' p}{\Sigma g' p'} \cdot \frac{\Sigma g p}{\Sigma g p'}}.$$

[1]) Vgl. Die Messung der Lebenshaltungskosten, Aufgaben und Praxis der Indexberechnung, Vierteljahrshefte zur Statistik des Deutschen Reichs 1937, 1. Heft, S. 149.

[2]) Vgl. Statistisches Jahrbuch für das Deutsche Reich 1936, S. 294, und Wirtschaft und Statistik (laufend).

Hierin bedeuten die Werte g' die Warenmengen zur Basiszeit und die Werte g die Warenmengen zur Messungszeit. Die Doppelformel stellt sich also dar als das geometrische Mittel aus den beiden Werten des Lebenshaltungsindex, die erhalten werden, wenn zur Basiszeit und zur Messungszeit verschiedene Gütermengen in Ansatz gebracht werden. Die genannten beiden Schwierigkeiten werden bei Anwendung der Doppelformel in der Weise behoben, daß man die g-Werte für die Güter, die zur Messungszeit nur noch in geringem Maße gebraucht werden, sehr niedrig und die g-Werte für die Güter, die überhaupt nicht mehr vorkommen, mit 0 in Ansatz bringt. Die Anwendung der Doppelformel erfordert wesentlich mehr Rechnung als die Anwendung der einfachen Indexformel.

Die zweite Möglichkeit zur Behebung der genannten Schwierigkeiten besteht darin, daß man eine Verkettung ausführt, indem man annimmt, daß sich die Preise der Waren, die zur Messungszeit nur noch in geringem Grade oder überhaupt nicht mehr in Betracht kommen, relativ in derselben Weise bewegen oder bewegt hätten wie die Preise der Waren, die an ihrer Stelle verwendet werden[1]).

[1]) Vgl. hierzu A. Jacobs, Die allgemeine Preisindexziffer im Dienste der Realwertrechnung. Allgemeines Statistisches Archiv, 23. Bd., S. 305 u. f. und L. v. Bortkiewicz, Zweck und Struktur einer Preisindexzahl. Nordisk Statistisk Tidskrift 1923, S. 369 und 1924, S. 208 und 409, und v. Bortkiewicz, Der gegenwärtige Stand des Problems der Geldwertmessung. Handwörterbuch der Staatswissenschaften, 4. Aufl., Bd. IV, 1926, S. 743 u. f.

Nicht bloß auf dem Gebiete der Lebenshaltungskosten, sondern auch auf anderen wichtigen Lebensgebieten werden Indexziffern berechnet, so für die landwirtschaftliche und industrielle Erzeugung, für die landwirtschaftlichen Verkaufsprodukte, für die Großhandelspreise, für die Fertigwarenpreise u. a. Im Rahmen der Statistik der Fertigwarenpreise hat H. Platzer eingehende Feststellungen über den Stand und die Ergebnisse der Indexberechnungen in den verschiedenen Ländern getroffen (H. Platzer, Zur Statistik der Fertigwarenpreise. Bulletin de l'Institut international de Statistique, tome XXVII, livraison 2, 1934, p. 452 u. f.).

Die Indexziffernmethode findet auch in der Statistik der Produktivität Anwendung. Die hier auftretenden Probleme hat F. Zahn einer eingehenden Untersuchung unterzogen. Zahn entwickelt, daß die volkswirtschaftliche Produktivität durch den Quotienten aus dem Index des Produktionsvolumens und dem Index der geleisteten Arbeitsstunden zu messen ist (F. Zahn, Statistik der Produktivität. Allgemeines Statistisches Archiv, 22. Bd., S. 530 u. f.).

Die Indexmethode wird nach Darlegungen von A. Zwick in der Gesundheitsstatistik zur Kennzeichnung des Gesundheitsniveaus herangezogen werden können, sobald die Medizin die hierfür erforderlichen Grundlagen geschaffen hat (A. Zwick, Gesundheitsstatistik, ein Beitrag zur Problemstellung. Allgemeines Statistisches Archiv, 23. Bd., S. 489 u. f.).

In der Literatur ist auch die Frage der Auswertung des Warenpreisindex für währungspolitische Zwecke erörtert worden. Es sei in diesem Zusammenhang auf die Untersuchungen von Ch. Lorenz hingewiesen (Ch. Lorenz, Die statistischen Grundlagen der Indexwährung. Allgemeines Statistisches Archiv, 22. Bd., 1932, S. 492 u. f.).

Anschließend sei noch bemerkt, daß es auch Gebiete gibt, auf denen das Arbeiten mit zusammengesetzten statistischen Zahlen nicht zum Ziele führt. So hat W. Grävell gezeigt, daß sich über die Rohstoffversorgung mittels Gesamtzahlen kein Urteil gewinnen läßt. Hier ist es notwendig, auf die einzelnen Bewegungen zurückzugreifen (W. Grävell, Menge-Wert-Volumen. Zur Deutung der Außenhandelsstatistik. Allgemeines Statistisches Archiv, 27. Bd., S. 1 u. f.).

55. Schärfere Methoden für die Berechnung von totalen Beziehungszahlen. Den Unterschied zwischen totalen und partiellen Beziehungszahlen wollen wir an dem Beispiel der Sterbeziffer behandeln. Die totale Sterbeziffer wird in der Weise berechnet, daß man die Gesamtzahl der Gestorbenen eines Kalenderjahres in Beziehung zur mittleren Bevölkerungszahl setzt. Die partielle Sterbeziffer (Sterblichkeitskoeffizient) erhält man, wenn man z. B. die Gestorbenen eines bestimmten Altersjahres in Beziehung zur mittleren Zahl der Personen dieses Altersjahres setzt. Die Beziehungszahlen dienen in erster Linie der Vornahme von statistischen Vergleichen in zeitlicher, räumlicher und sachlicher Hinsicht. Verwendet man z. B. die totale Sterbeziffer zur Untersuchung des Sterblichkeitsrückganges, so ist zu bedenken, daß die Sterbeziffer in hohem Grade von der Altersstruktur der Bevölkerungsgesamtheit abhängt. Weiter ist die totale Sterbeziffer auch abhängig von der Familienstandsgliederung der Bevölkerung, von der Verteilung der Bevölkerung nach Stadt und Land, der Rassenzugehörigkeit, der wirtschaftlichen und sozialen Gliederung der Bevölkerung. Diese Faktoren wirken störend ein bei der Vornahme von sterblichkeitsstatistischen Vergleichen mittels der totalen Sterbeziffer, weiter stören diese Faktoren auch die Durchführung von geburten- und heiratsstatistischen Vergleichen mittels der totalen Geburten- und totalen Eheschließungsziffer. Bei der totalen Sterbeziffer können außerdem noch Strukturunterschiede in bezug auf Geschlecht, Legitimität, klimatische, geographische und geologische Verhältnisse einen störenden Einfluß ausüben; bei der Berechnung von totalen Geburtenziffern sind als weitere Vergleichsstörungen Unterschiede hinsichtlich des Heiratsalters, der Ordnungszahl der Ehe, der Religionsgliederung und der Säuglingssterblichkeit in Betracht zu ziehen. Diese störenden Faktoren lassen sich, wie Žižek[1] ausführt, durch sachliche Differenzierung ausschalten. Die sachliche Differenzierung besteht in der Berechnung von partiellen Beziehungszahlen. Will man neben der sachlichen Differenzierung noch einen einfachen einheitlichen Vergleich mittels totaler Beziehungszahlen vornehmen, so stehen zwei Methoden zur Verfügung: die Standardisierungs- und die Tafelmethode.

Daß es unbedingt notwendig ist, die vergleichsstörenden Faktoren auszuschalten, geht aus der Tatsache hervor, daß auf Grund von Gefügeverschiedenheiten folgender Fall eintreten kann: die partiellen Beziehungszahlen liegen im Lande A höher als im Lande B, während die totale Beziehungszahl im Lande B einen höheren Wert aufweist als in A. Zur Auffindung der notwendigen und hinreichenden Bedingung für das Eintreten einer derartigen Umkehrung sei unter Verwendung des Beispiels der Sterbeziffer die Gesamtbevölkerung in zwei Altersgruppen aufgeteilt und die partielle Sterbeziffer der ersten Altersgruppe im Lande A mit k_1, im Lande B mit k_1' und die partielle Sterbeziffer der zweiten Altersgruppe im Lande A mit k_2, im Lande B mit k_2' bezeichnet. Der relative Anteil der Bevölkerung der ersten Altersgruppe an der Gesamtbevölkerung sei im Land A l_1 und im Land B l_1'. Für die zweite Altersgruppe seien die entsprechenden Quoten im Lande A l_2 und im Lande B l_2'. Die Gesamtsterbeziffer k bzw. k' berechnet sich für das Land A auf

$$k = l_1 k_1 + l_2 k_2$$

und für das Land B auf

$$k' = l_1' k_1' + l_2' k_2'.$$

[1]) F. Žižek, Der statistische Vergleich. Allgemeines Statistisches Archiv, 21. Band, 1931, S. 532.

Durch eine verhältnismäßig leichte Betrachtung findet man, daß der folgende Fall

$$k < k'; \; k_1 > k'_1; \; k_2 > k'_2$$

dann und nur dann eintritt, wenn folgende Ungleichung erfüllt ist:

$$l_1 \frac{k_2 - k_1}{k'_2 - k'_1} - \frac{k_2 - k'_2}{k'_2 - k'_1} < l'_1.$$

Zur Behebung dieses Widerspruchs und zur allgemeinen Ausschaltung der vergleichsstörenden Faktoren sei zunächst die Standardisierungsmethode herangezogen. Bei dieser Methode zerlegt man die Grund- oder Bezugsmasse (Nennermasse) im Hinblick auf die auszuschaltenden Merkmale in Teilmassen (bei der Sterbeziffer ist die Gesamtbevölkerung die Grundmasse). Die Anzahl der auf die einzelnen Teilmassen entfallenden Elemente (bei der Sterbeziffer Personen) sei a_0, a_1. . . . Ebenso zerlegt man die bezogene Masse (Zählermasse), die zur Grundmasse in Beziehung gesetzt werden soll, nach den gleichen Merkmalen in die entsprechenden Teilmassen (bei der Sterbeziffer ist die Gestorbenengesamtheit die bezogene Masse). Die Anzahl der Elemente in den Teilgesamtheiten der bezogenen Masse sei b_0, b_1, . . . Ferner führt man eine dritte statistische Masse, die Standardmasse, in die Betrachtung ein, die ebenfalls nach den gleichen Merkmalen in Teilgesamtheiten zerlegt wird (bei der Sterbeziffer ist die Standardmasse eine normal zusammengesetzte Bevölkerungsgesamtheit). Die Umfänge dieser Teilgesamtheiten seien a'_0, a'_1,... Schließlich konstruiert man eine vierte statistische Masse in der Weise, daß sie zur Standardmasse in demselben Verhältnis steht wie die bezogene Masse zur Bezugsmasse Für diese vierte Gesamtheit, die als die auf die Standardmasse bezogene Masse bezeichnet sei, erhält man ganz von selbst die entsprechenden Teilgesamtheiten, deren Umfänge mit b'_0, b'_1, . . . bezeichnet seien. Es ist

$$b'_0 = \frac{b_0}{a_0} a'_0, \quad b'_1 = \frac{b_1}{a_1} a'_1 \; \ldots$$

Der Gesamtumfang $\Sigma b'$ der auf die Standardmasse bezogenen Gesamtheit ist somit

$$\Sigma b' = \Sigma \frac{b}{a} a'. \tag{4}$$

Die Summenzahl $\Sigma b'$ setzt man in Beziehung zur Gesamtzahl der Standardmasse, und der sich hierbei ergebende Quotient ist die standardisierte Beziehungszahl α.

$$\alpha = \Sigma b' : \Sigma a'. \tag{5}$$

Die standardisierte Beziehungszahl ist vollkommen unabhängig von der Struktur der Bezugsmasse in bezug auf die Merkmale, die der Standardisierung zugrunde gelegt werden. Sie ist lediglich abhängig von der Struktur der Standardmasse. Hieraus folgt, daß alle auf ein und dieselbe Standardmasse standardisierten Beziehungszahlen vergleichbar sind.

Über die Vergleichbarkeit von Geburten- und Sterbeziffern hat F. Burgdörfer[1]) grundlegende Untersuchungen angestellt. Er betont mit Nachdruck, daß es gegen-

[1]) F. Burgdörfer, Volk ohne Jugend. 3. Aufl., Berlin 1937, S. 28.

wärtig nicht angängig ist, die gewöhnliche totale Geburtenziffer mit der gewöhnlichen totalen Sterbeziffer zu vergleichen, da infolge des gestörten Altersaufbaus die erstere Ziffer zu hoch und die letztere Ziffer zu niedrig liegt. Um diese beiden Verhältniszahlen vergleichbar zu machen, bereinigt sie Burgdörfer von dem störenden Einfluß der Altersgliederung. Hierbei kommt er zu dem Ergebnis, daß sich im Deutschen Reich für das Jahr 1925 die bereinigte Geburtenziffer auf 15,9 a. T. und die bereinigte Sterbeziffer auf 17,4 a. T. stellt. Hiernach beläuft sich das Geburtendefizit auf 1,5 a. T. Für 1933 lauten die entsprechenden Zahlen 11,7; 16,4 und 4,7[1]).

Im folgenden wollen wir eine mathematische Methode für die Ausschaltung störender Einflüsse entwickeln.

Für das Bevölkerungsproblem der Gegenwart ist die Standardisierung der Geburtenziffer von Wichtigkeit. Hierbei betrachtet man die Gesamtheit der weiblichen Personen im Alter von 15 bis 45 Jahren als die Bezugsmasse und die Gesamtheit der Lebendgeborenen als die bezogene Masse. Man gliedert zunächst die Gesamtheit der weiblichen Personen nach dem Alter x. Die hierbei entstehenden Teilgesamtheiten enthalten a_x Personen. Jede Teilgesamtheit zerlegt man weiter nach dem Familienstand y. Die sich auf diese Weise ergebenden Teilmassen 2. Ordnung umfassen a_{xy} Personen. Aus den letzteren Teilmassen 2. Ordnung wählt man die aus, bei denen das Merkmal y den Wert verheiratet aufweist und spaltet sie weiter nach der Ehedauer z in Teilgesamtheiten 3. Ordnung auf, deren Personenzahl a_{xyz} sei. Diese Teilgesamtheiten 3. Ordnung gliedert man weiter nach der Ordnungszahl der Ehe ξ in Teilgesamtheiten 4. Ordnung und bezeichnet deren Personenzahl mit $a_{xyz\xi}$. Der Index x durchläuft die Alterszahlenwerte 15 bis 44, y nimmt die vier Familienstandswerte ledig, verheiratet, verwitwet, geschieden an; z läuft von 0 bis etwa 29 und ξ beginnt mit 1.

In entsprechender Weise zerlegen wir die Gesamtheit der Lebendgeborenen nach dem Lebensalter, dem Familienstand, der Ehedauer und der Ehenzahl der Mütter und bezeichnen mit $b_{xyz\xi}$ die Zahl der Lebendgeborenen, deren Mütter im Alter von x bis $x+1$ Jahren im Familienstand y im Ehejahr z bis $z+1$ in der ξ-ten Ehe sich befinden.

Die Umfänge der nunmehr einzuführenden Standardbevölkerung seien allgemein mit $a'_{xyz\xi}$ bezeichnet. Daraus berechnen sich sofort die Umfänge $b'_{xyz\xi}$ der Teilgesamtheiten der auf die Standardbevölkerung bezogenen Lebendgeborenengesamtheit

$$b'_{xyz\xi} = \frac{b_{xyz\xi}}{a_{xyz\xi}} a'_{xyz\xi} .$$

Die Gesamtzahl der Lebendgeborenen der Standardbevölkerung läßt sich in der Form $\Sigma b'_{xyz\xi}$ und die der Personen der Standardbevölkerung in der Form $\Sigma a'_{xyz\xi}$ darstellen. Hieraus erhält man die standardisierte Geburtenziffer $\alpha_{xyz\xi}$ in der Form

$$\alpha_{xyz\xi} = \Sigma b'_{xyz\xi} : \Sigma a'_{xyz\xi} .$$

Als Standardbevölkerung kann man nach Auffassung von Körösy[2]) die Bevölkerung Schwedens wählen, weil diese nahezu eine normale mittlere Struktur

[1]) Sonderheft 15 zu Wirtschaft und Statistik, S. 78 und 79.

[2]) Körösy, Mortalitätskoeffizient und Mortalitätsindex. Bulletin de l'Institut international de Statistique, VI, 2, Rome 1892, S. 305.

aufweist. Nach L. March und Huber[1]) empfiehlt es sich, die Gesamtbevölkerung von 19 europäischen Staaten zur Zeit der Jahrhundertwende als Standardbevölkerung zu verwenden. Außerdem kommt für die Standardisierung nach dem Alter noch die stationäre Bevölkerung, die sich nach der Sterbetafelmethode ergibt, als Standardbevölkerung in Betracht.

Zur Erläuterung dieses Sachverhalts sei davon ausgegangen, daß in der Sterbetafel die Sterbenswahrscheinlichkeiten für die einzelnen Lebensjahre berechnet werden. Bezeichnet man mit $q_0, q_1, q_2, \ldots q_{\omega-1}$ die Sterbenswahrscheinlichkeiten des 1., 2., 3., . . ω-ten Lebensjahres, mit L_0 die Zahl der Lebendgeborenen eines Kalenderjahres und mit $L_1, L_2, \ldots L_\omega$ die Zahl derjenigen Personen, die von den L_0 Lebendgeborenen das 1., 2., 3., . . . ω-te Lebensjahr überleben, so ist

$$L_1 = L_0\,(1 - q_0)$$
$$L_2 = L_1\,(1 - q_1)$$
$$\cdot$$
$$L_\omega = L_{\omega-1}\,(1 - q_{\omega-1}).$$

Die Zahlenfolge $L_0, L_1, L_2, \ldots L_\omega$ bezeichnet man als die Überlebensordnung. Aus den Zahlen der Überlebensordnung lassen sich sofort die mittleren Lebendenzahlen für jedes Altersjahr berechnen. Für das $(i+1)$te Lebensjahr beträgt die mittlere Lebendenzahl

$$L'_i = \frac{L_i + L_{i+1}}{2} = L_i\,(1 - \tfrac{1}{2}\,q_i). \tag{6}$$

Diese mittleren Lebendenzahlen kennzeichnen die zu einem bestimmten Zeitpunkt in den einzelnen Altersjahren lebenden Personen derjenigen Bevölkerung, die jahraus, jahrein die gleiche Lebendgeborenenzahl, die gleichen Sterbenswahrscheinlichkeiten aufweist und die keine Zu- und Abwanderung erfährt. Eine so beschaffene Bevölkerung bezeichnet man als stationäre Bevölkerung und die mittleren Lebendenzahlen $L'_0, L'_1, \ldots L'_\omega$ kennzeichnen den Altersaufbau der stationären Bevölkerung. Der Umfang der stationären Bevölkerung sei P.

$$P = \Sigma\, L'.$$

Die Zahl der Sterbefälle in der stationären Bevölkerung ist gleich der Zahl der Lebendgeborenen, also gleich L_0. Mithin berechnet sich die Sterbeziffer für die stationäre Bevölkerung β auf

$$\beta = \frac{L_0}{P}.$$

Man bezeichnet diese Sterbeziffer als Tafelsterbeziffer.

Es läßt sich nun allgemein zeigen, daß die Tafelsterbeziffer gleich ist der nach dem Alter auf die stationäre Bevölkerung standardisierten Sterbeziffer. Zu diesem Zwecke gehen wir von der Formel (4) dieses Artikels aus und berechnen zunächst

[1]) Vgl. Annuaire International de Statistique, Bd. II, S. VIII.

für die einzelnen Altersjahre die Quotienten $\frac{b}{a}$, wobei mit Bezug auf ein bestimmtes Altersjahr a die mittlere Zahl der Lebenden und b die Zahl der Gestorbenen bedeuten. Der Quotient $\frac{b}{a}$ ist gleich dem Sterblichkeitskoeffizienten k. Die Größen a_0', a_1', a_2', . . , sind in unserem Falle gleich den mittleren Lebendenzahlen L_0', L_1', L_2', . . . L_ω'. Infolgedessen geht die allgemeine Formel (4) über in

$$\Sigma\, b' = \Sigma\, k\, L'.$$

Nach (5) stellt sich somit die standardisierte Sterbeziffer auf

$$\alpha = \frac{\Sigma\, k\, L'}{\Sigma\, L'}.$$

Nach Formel (3b) in Art. 53 können wir für k schreiben

$$\kappa = \frac{}{1-\frac{q}{2}}.$$

Somit ist

$$\Sigma\, k\, L' = \sum \frac{q\, L'}{1-\frac{q}{2}}.$$

Für L' setzen wir nach (6)

$$L' = L\left(1-\frac{q}{2}\right).$$

Folglich erhalten wir

$$\alpha = \frac{\sum \frac{q\, L\left(1-\frac{q}{2}\right)}{1-\frac{q}{2}}}{\Sigma\, L'} = \frac{\Sigma\, q\, L}{\Sigma\, L'} = \frac{\Sigma\, q\, L}{P}.$$

$\Sigma\, q\, L$ ist gleich der Summe der Gestorbenenzahlen aller Lebensjahre. Da in der stationären Bevölkerung die Gestorbenenzahl gleich der Lebendgeborenenzahl ist, so ist $\Sigma\, q\, L = L_0$, also

$$\alpha = \frac{L_0}{P}.$$

Damit ist unsere Behauptung bewiesen[1]).

Die im vorstehenden entwickelte Sterbetafeltheorie bezieht sich auf einfach abgestufte Tafeln (Abstufung nach dem Alter). In der Versicherungspraxis werden auch doppelt (nach Alter und Versicherungsdauer) abgestufte Tafeln benötigt. Die

[1]) Vgl. hierzu F. Burkhardt, Die Standardisierungs- und die Tafelmethode im Dienste der statistischen Praxis. Beiträge zur deutschen Statistik. Festgabe für Franz Žižek, 1936, S. 61 u. f.

Berechnung und Ausgleichung dieser Tafeln ist eingehend von G. Höckner[1], A. Abel[2]) und P. E. Böhmer[3]) behandelt worden. Der Böhmersche Ansatz schließt den Höcknerschen und den Abelschen Ansatz als Spezialfälle in sich. Die praktische Durchführung der Berechnung und Ausgleichung von Sterbetafeln ist von F. Burkhardt[4]) dargelegt worden.

Im Anschluß hieran sei noch die Frage erörtert, in welcher Weise die Standardisierung nach dem Familienstand durchzuführen ist. Nach statistischen Untersuchungen kann angenommen werden, daß die Familienstandsgliederung nach der Volkszählung 1910 als nahezu normal angesehen werden kann. Man kann beweisen, daß die Familienstandsquoten von 1910 ohne besondere Umrechnung auf die stationäre Bevölkerung übertragen werden können. Zu diesem Ende bezeichnen wir mit $F(x)$ die Zahl der in einem bestimmten Familienstand beim Alter von x Jahren lebenden Personen, die im Zeitintervall (t_1, t_2) von der Länge eines Jahres geboren wurden. Weiter stellen wir mit $\Phi(x)\,dx$ den Mehrabgang von Personen aus der Geburtszeit (t_1, t_2) im Altersintervall $(x, x+dx)$ dar.

$$\Phi(x) = -\frac{dF(x)}{dx}$$

Somit ist der Mehrabgang von Personen aus der Geburtszeit $(t_1+t,\ t_1+t+dt)$ im Altersintervall $(x, x+dx)$ gleich $\Phi(x)\,dt\,dx$. Der gesamte Mehrabgang von Personen aus der Geburtszeit (t_1, t_2) vom Beginn des Altersjahres (x_1, x_2) bis zum Volkszählungszeitpunkt $\tau = t_1 + x_2 = t_2 + x_1$ stellt sich also auf

$$\int_0^1 \int_{x_1}^{x_2-t} \Phi(x)\,dt\,dx,$$

da t im Geburtszeitintervall (t_1, t_2) von 0 bis 1 läuft. Infolgedessen ergibt sich für die am Volkszählungszeitpunkt τ festgestellte Zahl der Personen aus der Geburtszeit (t_1, t_2) im Altersjahr (x_1, x_2)

$$V_\tau = F(x_1) - \int_0^1 \int_{x_1}^{x_2-t} \Phi(x)\,dt\,dx.$$

Hieraus folgt weiter

$$V_\tau = F(x_1) - \int [F(x_1) - F(x_2-t)]\,dt$$

$$V_\tau = F(x_1) - F(x_1)[t]_0^1 + \int_0^1 F(x_2-t)\,dt.$$

[1]) G. Höckner, Änderung der Rechnungsgrundlagen, Leipzig 1907, S. 27 u. f.

[2]) A. Abel, Sterbetafeln 1926, Heft 40 der Veröffentlichungen des Deutschen Vereins für Versicherungswissenschaft, S. 18 u. f.

[3]) P. E. Böhmer, Betrachtungen über die deutschen Sterbetafeln 1926, Zeitschrift für die gesamte Versicherungswissenschaft 1927, S. 176 u. f.

[4]) F. Burkhardt, Die neue Sterbetafel für die Gesamtbevölkerung Sachsens im Anschluß an die Volkszählung am 16. Juni 1925. Zeitschrift des Sächsischen Statistischen Landesamtes 1928/29, S. 103 u. f.

Mit Hilfe der Substitution

$$x = x_2 - t$$

findet man, da $x_2 - x_1 = 1$ ist,

$$V_\tau = \int^{x_2} F(x)\,dx.$$

Die Volkszählungszahl V_τ ist somit gleich der im Altersjahr (x_1, x_2) verlebten Jahre und folglich auch gleich der mittleren Personenzahl dieses Altersjahres. Hieraus ergibt sich die Schlußfolgerung, daß die bei der Volkszählung ermittelten Familienstandsquoten auf die stationäre Bevölkerung übertragen werden können.

56. Die Berechnung von Beziehungszahlen erfordert dann besondere Überlegungen, wenn sich an der Nenner- oder Bezugsmasse nicht bloß die Veränderung vollzieht, deren Ereignisfälle in der Zählermasse erscheinen, sondern wenn außerdem Veränderungen einer anderen Art eintreten. Dieser Fall liegt z. B. vor bei der Messung der Sterblichkeit des ersten Lebensjahres, wenn es gilt, die Messung getrennt für die Ehelichen und Unehelichen vorzunehmen. Zur Lösung dieses Problems führen wir verschiedene biometrische Funktionen ein, und zwar zunächst die Überlebensfunktion $f(x)$. Es bezeichne $1000\,f(x)$ die Zahl der Personen, die sich von 1000 Lebendgeborenen beim Alter von x Jahren noch am Leben befinden. Aus der Funktion $f(x)$ leiten wir die Absterbefunktion $\varphi(x) = -\frac{df(x)}{dx}$ her. Die Größe $\varphi(x)\,dx$ gibt an, wie groß für die Lebendgeborenen die Wahrscheinlichkeit ist, im Alter von x bis $x + dx$ Jahren zu sterben. Weiter führen wir die Bestandsfunktion der Unehelichen $g(y)$ ein. Es bedeutet $1000\,g(y)$ die Zahl der Kinder, die sich von 1000 unehelich Lebendgeborenen beim Alter von y Jahren noch im unehelichen Stand befinden, wenn der Abgang nur durch Legitimation, nicht auch durch Tod erfolgt. Aus der Funktion $g(y)$ leiten wir die Legitimierungsfunktion $\psi(y) = -\frac{dg(y)}{dy}$ her. Die Größe $\psi(y)\,dy$ kennzeichnet die Wahrscheinlichkeit, die die unehelich Lebendgeborenen haben, im Alter von y bis $y + dy$ Jahren legitimiert zu werden. Außerdem bezeichnet V_e bzw. V_u die Zahl der ehelich bzw. unehelich Lebendgeborenen eines Kalenderjahres mit den Grenzen $t = 0$ und $t = 1$. Bei gleichmäßiger Verteilung der Lebendgeburtsfälle über das ganze Kalenderjahr kann die Zahl der ehelich Lebendgeborenen im Zeitelement dt in der Form $V_e\,dt$ angesetzt werden. Von den ehelich Lebendgeborenen des Geburtskalenderjahres sterben bis zum Ende des Geburtskalenderjahres

$$\int_0^1 \int_0^{1-t} V_e\,\varphi(x)\,dt\,dx.$$

Bezeichnet man mit γ_e bzw. γ_u die Sterbenswahrscheinlichkeiten der Ehelichen bzw. der Unehelichen im ersten Lebensjahr, so ist die Zahl der Sterbefälle der Ehelichen im ersten Lebensjahr durch den Ausdruck $V_e\,\gamma_e$ darstellbar. Nach statistischen Erfahrungen sterben bei gleichmäßiger Verteilung der Geburtenfälle von der Gesamtheit der aus der Geburtenmasse eines Kalenderjahres im ersten Lebens-

jahr Absterbenden infolge der größeren Sterblichkeit in den ersten Lebensmonaten $\frac{7}{10}$ im Geburtskalenderjahr und $\frac{3}{10}$ im folgenden Kalenderjahr. Somit gilt die Gleichung

$$\int_0^1 \int_0^{1-t} V_e \, \varphi(x) \, dt \, dx = 0{,}7 \; V_e \gamma_e .$$

Um die Gesamtzahl der im Geburtskalenderjahr ehelich absterbenden Kinder zu erhalten, ist zu dem obenstehenden Ausdruck noch die Zahl der im Geburtskalenderjahr absterbenden legitimierten Kinder zu addieren. Unter Zuhilfenahme der eingeführten biometrischen Funktion läßt sich diese letztere Gestorbenenzahl in folgender Weise darstellen:

$$\int_0^1 \int_0^{1-t} \int_y^{1-t} V_u \, \psi(y) \, \varphi(x) \, dt \, dy \, dx .$$

In Anbetracht der kleinen Zahlen der Legitimationen und der legitimiert Gestorbenen gegenüber den aus der ehelichen Geburtenmasse Abgestorbenen kann an Stelle der Legitimierungsfunktion $\psi(y)$ und der Absterbefunktion $\varphi(x)$ mit der Legitimierungswahrscheinlichkeit ψ und der Sterbenswahrscheinlichkeit γ_e gerechnet werden. Unter Benutzung dieser vereinfachenden Annahmen ergibt sich für das dreifache Integral der Wert

$$\tfrac{1}{6} \cdot V_u \, \psi \, \gamma_e .$$

Das Produkt $V_u \psi$ wollen wir gleich $2\,E$ setzen, wobei wir mit E die Zahl der im Geburtsjahr legitimierten Kinder bezeichnen, denn die unehelich Lebendgeborenen haben im Geburtskalenderjahr im Mittel ein halbes Jahr die Möglichkeit, legitimiert zu werden. Die Gesamtzahl M_e der aus der Geburtenmasse des Kalenderjahres stammenden und im gleichen Kalenderjahr ehelich absterbenden Kinder ist also gleich

$$M_e = 0{,}7 \; V_e \gamma_e + \tfrac{1}{3} \, E \gamma_e .$$

Hieraus folgt:

$$\gamma_e = \frac{M_e}{0{,}7 \; V_e + \frac{1}{3} E} .$$

Für die Sterbenswahrscheinlichkeit der Unehelichen ergibt sich durch die entsprechende Überlegung der folgende Ausdruck

$$\gamma_u = \frac{M_u}{0{,}7 \; V_u - \frac{1}{3} E} .$$

Differenziert man die Sterbenswahrscheinlichkeiten γ_e und γ_u nach dem Geschlecht, so erhält man die folgenden vier Sterbenswahrscheinlichkeiten: γ_{me} (Sterbenswahrscheinlichkeit der ehelichen Knaben), γ_{we} (Sterbenswahrscheinlichkeit der ehelichen Mädchen), γ_{mu} und γ_{wu}. Nach der sächsischen amtlichen Statistik berechnen sich diese vier Sterbenswahrscheinlichkeiten für das Jahr 1934 auf:

$$\gamma_{me} = 5{,}74 \qquad \gamma_{we} = 4{,}51$$
$$\gamma_{mu} = 9{,}56 \qquad \gamma_{wu} = 7{,}58$$

Der Quotient $\frac{\gamma_{mu}}{\gamma_{me}}$ kennzeichnet die Übersterblichkeit der Unehelichen bei den Knaben und der Quotient $\frac{\gamma_{wu}}{\gamma_{we}}$ die der Unehelichen bei den Mädchen. Durch Vergleichung der beiden Quotienten kommt man zu der Ungleichung

$$\frac{\gamma_{mu}}{\gamma_{me}} < \frac{\gamma_{wu}}{\gamma_{we}}.$$

Diese Ungleichung besagt, daß die Übersterblichkeit der Unehelichen bei den Mädchen größer ist als bei den Knaben. Die Mädchen leiden somit unter den Unbilden der Unehelichkeit stärker als die Knaben. Man kann aus dieser Ungleichung die Annahme herleiten, daß das weibliche Geschlecht auf die äußeren Verhältnisse in stärkerem Maße reagiert als das männliche.

Im Vorstehenden ist die Sterbenswahrscheinlichkeit der Ehelichen und Unehelichen im ersten Lebensjahr auf Grund der Sterbefälle im Geburtskalenderjahr berechnet worden. In der praktischen Statistik wird man auch häufig vor die Aufgabe gestellt, diese Sterbenswahrscheinlichkeiten auf Grund der Sterbefälle im ganzen ersten Lebensjahr zu bestimmen. Zu diesem Ende bringt man die Zahl der Sterbefälle von ehelich Lebendgeborenen im ersten Lebensjahr auf die Form

$$V_e \int \varphi(x)\, dx = V_e \gamma_e.$$

Zu dieser Gestorbenengesamtheit treten noch diejenigen ehelich gestorbenen Kinder hinzu, die unehelich geboren, aber legitimiert wurden. Die Zahl dieser letzteren Sterbefälle läßt sich folgendermaßen darstellen:

$$\int \int V_u \psi(y) \varphi(x)\, dy\, dx.$$

Wir setzen wiederum an Stelle von $\psi(y)$ ψ und von $\varphi(x)$ γ_e und erhalten für das Doppelintegral

$$\tfrac{1}{2} V_u \psi \gamma_e.$$

Für die Gesamtheit der im ersten Lebensjahr ehelich Gestorbenen M'_e gewinnen wir somit den Ausdruck

$$M'_e = V_e \gamma_e + \tfrac{1}{2} V_u \psi \gamma_e.$$

Für das Produkt $V_u \psi$ setzen wir E', wobei E' die Zahl der Legitimierten im ersten Lebensjahr bedeutet. Folglich ist

$$\gamma_e = \frac{M'_e}{V_e + \frac{1}{2} E'}.$$

Für die Sterbenswahrscheinlichkeit γ_u der Unehelichen ergibt sich

$$\gamma_u = \frac{M'_u}{V_u - \frac{1}{2} E'}.$$

Nach der sächsischen amtlichen Statistik ist im Jahre 1934

$$\gamma_{me} = 5{,}72 \qquad \gamma_{we} = 4{,}50$$
$$\gamma_{mu} = 9{,}53 \qquad \gamma_{wu} = 7{,}56.$$

Es gilt auch auf Grund dieser Zahlenwerte die Ungleichung

$$\frac{\gamma_{mu}}{\gamma_{me}} < \frac{\gamma_{wu}}{\gamma_{we}}.$$

57. Vergleichung von Abgangswahrscheinlichkeiten. Die im vorstehenden Artikel angestellten Betrachtungen lassen sich einen Schritt weiterführen, wenn man sich für die betrachteten Kollektive (z. B. Gesamtheit der unehelich Lebendgeborenen) die Frage der Vergleichung von Abgangswahrscheinlichkeiten vorlegt. Es seien die Umfänge eines Kollektivs zu zwei verschiedenen Zeitpunkten A_1 und A_2. Jeden Zeitpunkt betrachtet man als Anfang eines einjährigen Zeitraums. Im ersten einjährigen Zeitraum betrage die Zahl der Abgänge der ersten Art (z. B. Abgänge durch Legitimation) B_1 und im zweiten B_2. Weiter sei die Zahl der Abgänge der zweiten Art (z. B. Abgänge durch Tod) im ersten Jahr C_1, im zweiten C_2. Die Zahl der Verbleibenden (z. B. nicht legitimierte, lebende uneheliche Kinder) sei bezeichnet für das Ende des ersten Jahres mit D_1 und für das Ende des zweiten Jahres mit D_2. Führen wir für die Abgangswahrscheinlichkeit für die Abgänge der ersten Art im ersten Jahr δ_1 und im zweiten Jahr δ_2 ein, so läßt sich die Zahl der Abgänge der ersten Art bei gleichmäßiger Verteilung der Abgänge für beide Jahre unter Weglassung der Indizes durch den folgenden Ausdruck darstellen

$$B = A\,\delta - \delta \int_0^1 C(1-t)\,dt.$$

Das Integral gibt die Summe der Jahre an, die für die einzelnen abgehenden Elemente nach deren Abgang auf die zweite Art bis zum Ende des Beobachtungsjahres verstreichen. (Gehen z. B. auf die zweite Art im Laufe des Beobachtungsjahres $C = 50\,000$ Elemente ab, so ist das Integral gleich 25000.) Sehen wir von der Voraussetzung der gleichmäßigen Verteilung der Abgänge auf die zweite Art ab, so können wir allgemein schreiben

$$B = A\,\delta - \delta\,\Theta\,C.$$

Hier bedeutet Θ für die Abgänge auf die zweite Art das Verhältnis der Zahl der Jahre, die nach dem Abgang bis zum Ende des Beobachtungsjahres verstreichen, zur Zahl der Abgänge. Somit erhalten wir für die Abgangswahrscheinlichkeit der Abgänge auf die erste Art die Formel

$$\delta = \frac{B}{A - \Theta\,C}.$$

Erfolgen die Abgänge auf die zweite Art gleichmäßig im Beobachtungsjahr, so ist $\Theta = \frac{1}{2}$. Konzentrieren sich die Abgänge auf die zweite Art in der ersten Hälfte des Beobachtungsjahres, so ist $\Theta > \frac{1}{2}$. Bei Konzentration in der zweiten Hälfte

des Beobachtungsjahres ist $\Theta < \frac{1}{2}$. Erfolgen sämtliche Abgänge auf die zweite Art am Anfang oder am Ende des Beobachtungsjahres, so ist $\Theta = 1$ bzw. $\Theta = 0$.

Die exakte Bestimmung von Θ ist mittels der Tafelmethode unter Berechnung der verlebten Zeit vorzunehmen. Diese oft sehr mühsame Berechnung kann jedoch in gewissen Fällen umgangen werden, wenn es nur darauf ankommt, die Abgangswahrscheinlichkeit für verschiedene Beobachtungsjahre (entsprechendes gilt auch für verschiedene räumliche Gebiete) zu vergleichen. Zur Darlegung dieser Verhältnisse gehen wir von den beiden Abgangswahrscheinlichkeiten δ_1 und δ_2 aus:

$$\delta_1 = \frac{B_1}{A_1 - \Theta C_1}, \quad \delta_2 = \frac{B_2}{A_1 - \Theta C_2}.$$

Wir haben hier die Anfangsbestandzahlen zu Beginn der beiden Beobachtungsjahre gleich angenommen. Ist dies in Wirklichkeit nicht der Fall, so können wir die Abgangszahlen auf den gleichen Anfangsbestand umrechnen. Die Annahme eines gleichen Anfangsbestands bedeutet also keine Einschränkung. Wir wollen also für die folgenden Betrachtungen ansetzen:

$$B_1 + C_1 + D_1 = B_2 + C_2 + D_2. \tag{7}$$

Zur leichteren Durchführung der folgenden Betrachtungen führen wir den Begriff der Mindestzahl der Abgangsfähigen $(B_1 + D_1)$ bzw. $(B_2 + D_2)$ und den Begriff der vollen Zahl der Abgangsfähigen $(B_1 + D_1 + [1 - \Theta]\, C_1)$ bzw. $(B_2 + D_2 + [1 - \Theta]\, C_2)$ ein. Hierzu sei mit Bezug auf das Beispiel der Abgänge durch Legitimation und Tod bemerkt, daß sowohl die Legitimierten als auch die im unehelichen Stande Verbleibenden als legitimationsfähig zu betrachten sind. Ihre Summe ergibt die Mindestzahl der Legitimationsfähigen. Zu dieser Mindestzahl kommt noch ein Teil der Gestorbenen hinzu, für die ebenfalls die Möglichkeit der Legitimation bestand. Ist für den zweiten Beobachtungszeitraum die Zahl der Mindestabgangsfähigen größer als für den ersten Beobachtungszeitraum, gilt also die Gleichung

$$B_2 + D_2 > B_1 + D_1, \tag{8}$$

so ist die volle Zahl der Abgangsfähigen für den zweiten Zeitraum ebenfalls größer als für den ersten Zeitraum, d. h. es ist

$$B_2 + D_2 + \Theta C_2 > B_1 + D_1 + \Theta C_1. \tag{9}$$

Weiter läßt sich allgemein zeigen, daß bei Bestehen der beiden letzten Ungleichungen die weitere Ungleichung existiert

$$\frac{B_2 + D_2}{B_1 + D_1} > \frac{B_2 + D_2 + \Theta C_2}{B_1 + D_1 + \Theta C_1}. \tag{10}$$

Die Richtigkeit dieser neuen Ungleichung sieht man sofort ein, wenn man bedenkt, daß infolge der Gleichung (7) bei Bestehen der Ungleichungen (8) und (9) $C_2 < C_1$ ist. Man kann also sagen: Wenn die Mindestzahl der Abgangsfähigen von einem Zeitraum zum andern zunimmt, so nimmt auch die volle Zahl der Abgangsfähigen

zu, aber in geringerem Grade. Hieraus folgt sofort: Ist die relative Zunahme der Zahl der Abgänge der betrachteten Art von einem ersten Zeitraum zu einem zweiten größer als die relative Zunahme der Mindestzahl der Abgangsfähigen, so ist die Abgangswahrscheinlichkeit für den zweiten Zeitraum größer als für den ersten. Die Richtigkeit dieser letzteren Schlußfolgerung kann man sofort in folgender Weise darlegen. Nach der Annahme ist

$$\frac{B_2}{B_1} > \frac{B_2 + D_2}{B_1 + D_1}.$$

Mit Bezug auf (10) ergibt sich dann sofort:

$$\frac{B_2}{B_2 + D_2 + (1-\Theta)\,C_2} > \frac{B_1}{B_1 + D_1 + (1-\Theta)\,C_1} \qquad (11)$$

Da

$$B_1 + D_1 + (1-\Theta)\,C_1 = A_1 - \Theta\,C_1$$

und

$$B_2 + D_2 + (1-\Theta)\,C_2 = A_1 - \Theta\,C_2,$$

so folgt aus (11)

$$\delta_2 > \delta_1.$$

Hierzu sei ein Beispiel aus der sächsischen Legitimationsstatistik angeführt. Im Zeitraum 1904 bis 1908 wurden in Sachsen von 100 unehelich Lebendgeborenen 31,04 legitimiert und 36,04 starben. Im Zeitraum 1909 bis 1913 betrugen die entsprechenden Zahlen 35,88 und 26,72. Somit stellte sich die Mindestzahl der Legitimationsfähigen im ersten Zeitraum auf 63,96 und im zweiten Zeitraum auf 73,28. Somit lag die letztere Zahl im zweiten Zeitraum um 14,75 % höher als im ersten. Dasselbe gilt in verstärktem Maße für die Prozentzahl der Legitimierten. Für sie berechnet sich die Steigerungszahl auf 15,59. Somit kann man vollkommen exakt die Schlußfolgerung ziehen, daß die Legitimationshäufigkeit im Zeitraum 1909 bis 1913, bezogen auf die 1909 unehelich Geborenen, höher lag als die Legitimationshäufigkeit im Zeitraum 1904 bis 1908, bezogen auf die 1904 unehelich Lebendgeborenen.

Zur Ergründung dieser Zusammenhänge ist auch die Tafelmethode angewandt worden. In dieser Richtung liegen Untersuchungen von W. Winkler[1]) und H. Cl. Nybølle[2]) vor. Nach neueren Forschungen, die Winkler dem Internationalen Kongreß für Bevölkerungswissenschaft in Paris 1937 vorlegte, schieden nach der für den Durchschnitt der Jahre 1929 bis 1934 berechneten Abgangsordnung in Österreich 39 ‰ der männlichen und 37 ‰ der weiblichen unehelich Lebendgeborenen bis zum Eintritt in das siebente Lebensjahr aus, wobei auf die Gestorbenen, bzw. die Legitimierten entfielen: beim männlichen Geschlecht 15,8 ‰, bzw. 23,2 ‰ und beim weiblichen Geschlecht 12,9 ‰, bzw. 23,8 ‰.

[1]) W. Winkler, Die statistischen Verhältniszahlen, Leipzig und Wien 1923, S. 82 u. f.

[2]) H. Cl. Nybølle, Aegtebørns og Uaegtebørns Dodelighed. Nationalokonomisk Tidsskrift 1923, 2. Heft, S. 114 u. f.

§ 4. Streuungsmaße.

58. Wenn man von der Streuung oder Dispersion eines Kollektivs und der hinter ihm stehenden Materie spricht, so denkt man dabei an zwei Umstände: an die Ausbreitung der Argumentwerte und an ihre stellenweise Häufung.

Die Ausbreitung wird gekennzeichnet durch den kleinsten und größten vorkommenden Argumentwert; beider Unterschied nennt man die Variationsweite oder -breite.

Es liegt die Vermutung nahe, daß zwischen Ausbreitung und Häufung eine funktionelle Abhängigkeit bestehe in dem Sinne, daß bei stärkerer Häufung, also bei stärkerem Zusammenrücken der Argumentwerte an einer bestimmten Stelle, die Variationsweite abnehmen werde und umgekehrt.

Wenn auch im allgemeinen ein solcher Sachverhalt zutrifft, so läßt sich von ihm doch kein rechnerischer Gebrauch machen, und zwar schon deshalb nicht, weil die Variationsweite sich bei zunehmendem Umfang unstetig ändert, während die Häufung dabei beständige Fortschritte macht. Es kann nämlich geschehen, daß selbst bei beträchtlicher Umfangsvermehrung die extremen Argumentwerte dieselben bleiben und wiederum, daß bei geringer Vermehrung des Umfangs eine plötzliche und beträchtliche Zunahme der Variationsbreite sich einstellt, indem ein erheblich kleinerer oder größerer Argumentwert als die bisherigen zur Beobachtung kommt.

Mit der Angabe der Variationsweite allein ist also nichts Entscheidendes zur Kennzeichnung der Streuung getan. Wenn man sie trotzdem anführt, so geschieht es, weil es ein Interesse hat, die extremen Argumentwerte innerhalb eines Kollektivs kennen zu lernen.

Sowie man sich bei einem Kollektiv zu der beobachteten Verteilung eine ideelle hinzudenkt, die gewissermaßen die Grenze der Verteilung bei unbeschränkt wachsendem Umfang darstellt, so kann man sich auch zu den beobachteten Extremen ideelle hinzudenken, Werte also, welche das Argument niemals, auch bei noch so großer Vermehrung des Umfangs, unter-, beziehungsweise überschreitet. Doch ist es unsicher, aus der ideellen Häufigkeitskurve, die man der beobachteten Verteilung angepaßt hat, auf die ideellen Extreme schließen zu wollen; eine kleine Abänderung an der Kurve kann sie stark verschieben. Man wird also, wo nicht aus der Natur der Sache ein Anhalt zu gewinnen ist, auf die Bestimmung der ideellen Extreme verzichten müssen.

Jedes brauchbare Streuungsmaß wird sich auf die Gesamtheit der beobachteten Argumentwerte und ihre Häufigkeiten stützen müssen, wird auch im übrigen ähnliche Forderungen zu erfüllen haben, wie sie bezüglich der Mittelwerte gestellt worden sind, also Bestimmtheit und möglichst leichte Erfaßbarkeit; erwünscht ist auch eine einfache bequeme Berechnungsweise; algebraische Vorteile erhöhen den Wert eines solchen Maßes.

Als Vorbild hat hier die Fehlertheorie gedient. Das verrät schon die Namengebung; häufig werden die Formeln der Fehlertheorie ohne jede Abänderung auf das neue Gebiet übertragen, obwohl die Sachlage nicht die gleiche ist. Selbst der Begriff der Genauigkeit wird aus der Fehlertheorie in die Statistik herübergenommen, wiewohl das Streuungsmaß ein ganz anderes Moment zum Ausdruck bringt: Veränderlichkeit, Beständigkeit.

In den folgenden Artikeln werden die wichtigsten und gebräuchlichsten Streuungsmaße kritisch besprochen und die Methoden ihrer Berechnung auseinandergesetzt.

59. Die weiteste Verbreitung hat unter den Streuungsmaßen die mittlere Abweichung erlangt. Darunter wird die Quadratwurzel aus dem Durchschnitt der Quadrate der Abweichungen der einzelnen Argumentwerte von ihrem arithmetischen Mittel verstanden.

Ist X ein einzelner Argumentwert, z seine Häufigkeit, M das arithmetische Mittel, $X - M = \delta$ die Abweichung, N der Umfang des Kollektivs, so ist die mittlere Abweichung μ der X von M gleich dem quadratischen Mittel aus den δ-Werten (vgl. Art. 50)

$$\mu = \sqrt{\frac{1}{N} \Sigma (z \delta^2)}. \tag{1}$$

Neben der oben gewählten Benennung sind auch andere gebräuchlich: so mittlere quadratische Abweichung, um auf das Quadrieren der Einzelabweichungen hinzuweisen; ferner Streuung[1]) schlechtweg, so daß dieses Wort in doppeltem Sinne gebraucht wird, zur Bezeichnung der allgemeinen Erscheinung, die wir als Streuung erklärt haben, und eines besondern Maßes ihrer Intensität, schließlich Standardabweichung (standard-deviation) nach einem Vorschlage K. Pearsons[2]).

Die mittlere quadratische Abweichung, das quadratische Mittel und ebenso das arithmetische Mittel lassen sich in übersichtlicher Weise mittels des aus der theoretischen Mechanik entnommenen Begriffs des Moments darstellen. Sind X_1, $X_2, \ldots X_n$ die Argumentwerte und $z_1, z_2, \ldots z_n$ ihre Häufigkeiten, so lassen sich die folgenden Momente, bezogen auf den Anfangspunkt 0, bilden:

$$M_1 = \frac{1}{N} \sum_{i=1}^{n} z_i X_i, \quad M_2 = \frac{1}{N} \sum_{i=1}^{n} z_i X_i^2, \quad M_3 = \frac{1}{N} \sum_{i=1} z_i X_i^3 \ldots$$

Man bezeichnet M_1, bzw. M_2, bzw. $M_3, \ldots$ als Moment ersten, bzw. zweiten, bzw. dritten ... Grades, bezogen auf den Anfangspunkt 0. Das Moment ersten Grades M_1 ist gleich dem arithmetischen Mittel M.

Wählt man bei der Momentbildung als Bezugspunkt das arithmetische Mittel M, so erhält man die folgenden Momente:

$$M_1' = \frac{1}{N} \sum_{i=1}^{n} z_i (X_i - M), \quad M_2' = \frac{1}{N} \sum_{i=1}^{n} z_i (X_i - M)^2, \quad M_3' = \frac{1}{N} \sum_{i=1}^{n} z_i (X_i - M)^3 \ldots$$

Das Moment ersten Grades M_1', bezogen auf das arithmetische Mittel M, hat den Wert 0; das Moment zweiten Grades M_2', bezogen auf das arithmetische Mittel M, ist gleich der mittleren quadratischen Abweichung.

Schon darin, daß sich die mittlere Abweichung an den wertvollsten Mittelwert, das arithmetische Mittel, anlehnt, liegt ein Vorzug derselben und gibt ihr eine auszeichnende Eigenschaft. Ließe man nämlich den Ausgangswert U, von welchem aus man die Einzelabweichungen messen will, unbestimmt, und bezeichnete die Abweichung $X - U$ mit ε, so würden zu verschiedenen U verschiedene $\Sigma (\varepsilon^2)$

[1]) H. Bruns, Wahrscheinlichkeitsrechnung und Kollektivmaßlehre. Leipzig und Berlin 1906, S. 119.

[2]) Phil. Trans. Roy. Soc., A, vol. 185 (1894), p. 80.

gehören; welche Rolle spielt nun M unter den U? Es führt zur kleinsten Summe der Abweichungsquadrate. In der Tat wird

$$\Sigma(\varepsilon^2) = (X_1 - U)^2 + (X_2 - U)^2 + \ldots + (X_N - U)^2$$

ein Minimum, wenn

$$\frac{d\,\Sigma(\varepsilon^2)}{d\,U} = -2\,[X_1 + X_2 + \ldots + X_N - N\,U] = 0,$$

also $U = \frac{1}{N}\Sigma(X) = M$ wird.

Aber auch die Abweichungen selbst, wenn sie sich auf das arithmetische Mittel beziehen, haben eine Eigenschaft, die sich bei weiterer Verfolgung der Theorie als förderlich erweist: Ihre Summe ist Null. Denn

$$\Sigma(\delta) = (X_1 - M) + (X_2 - M) + \ldots + (X_N - M) = \Sigma(X) - NM = 0 \qquad (2)$$

kraft der Definition des arithmetischen Mittels.

Zwischen dem arithmetischen Mittel und der mittleren Abweichung besteht eine Beziehung, von der man mit Nutzen Gebrauch machen kann: beider Quadratsumme ist das arithmetische Mittel der Quadrate der Einzelwerte von X. Es ist nämlich

$$X = M + \delta$$
$$X^2 = M^2 + 2\,M\delta + \delta^2,$$

infolgedessen und weil $\Sigma\,\delta = 0$,

$$\Sigma\,X^2 = NM^2 + \Sigma\,\delta^2,$$

woraus tatsächlich

$$M^2 + \mu^2 = \frac{1}{N}\Sigma\,X^2$$

folgt.

An sich böte die Berechnung von μ keine Schwierigkeit. Nachdem man M bestimmt hat, subtrahiert man es von allen statistisch erhobenen X, quadriert die Differenzen, nimmt ihren Durchschnitt unter Berücksichtigung der Häufigkeiten und zieht daraus die Quadratwurzel. Aber schon bei einem mäßigen Umfang und insbesondere dann, wenn M, wie das in der Regel der Fall, mehrere Dezimalstellen hat, erweist sich dieses direkte Verfahren als beschwerlich.

Zu einer wesentlichen Vereinfachung führt die folgende Überlegung. Bestimmt man einmal die Abweichung ε von einem beliebigen Wert U, ein zweitesmal die Abweichung δ von M, so besteht zwischen ihnen die Beziehung

$$\varepsilon - \delta = X - U - (X - M) = M - U = \eta,$$

aus der

$$\Sigma(\varepsilon^2) = \Sigma(\delta^2) + 2\,\eta\,\Sigma(\delta) + N\eta^2,$$

also wegen (2)

$$\Sigma(\varepsilon^2) = \Sigma(\delta^2) + N\eta^2$$

folgt; daraus ergibt sich, wenn man den Durchschnitt der ε^2 mit m^2 bezeichnet,

$$m^2 = \mu^2 + \eta^2 \qquad (3)$$

und umgekehrt

$$\mu^2 = m^2 - \eta^2. \qquad (4)$$

Hat man also m^2 berechnet, so ergibt sich daraus μ^2 durch Subtraktion von η^2. Man hat es nun in der Hand, U so zu wählen, daß die ε eine möglichst bequeme Rechnung zulassen.

Bei einem unstetigen Kollektiv, bei dem die Argumentwerte X in der Regel aufeinanderfolgende ganze Zahlen sind, wird man für U zweckmäßig einen dieser Werte wählen, wodurch auch die ε ganze Zahlen werden. Von der Berechnung des arithmetischen Mittels her ist η bekannt, folglich alles zur Berechnung von μ gegeben.

Die zweckmäßige Anlage der Rechnung ist aus der folgenden Tabelle zu ersehen, der das in Art. 33 besprochene Material aus Zählungen an Eschen zugrundeliegt; es betrifft die Verteilung der Fieder nach der Anzahl der Blättchen, die sie tragen.

Besser als durch M und μ ist die vorliegende botanische Erscheinung durch folgende Angaben beschrieben:

Anteil der Fieder mit ungerader Blättchenzahl .... 87,2 %
Anteil der Fieder mit gerader Blättchenzahl .... 12,8 %
Dominierende Blättchenzahlen 9 und 11 .... 65,7 %
davon entfallen 47,6 % auf 9 und 52,4 % auf 11
Am meisten bevorzugte Blättchenzahl 11 .... 34,5 %

Berechnung von μ.

X	z	ε	$z\varepsilon$	$z\varepsilon^2$
3	8	— 6	48	288
4	5	— 5	25	125
5	142	— 4	568	2272
6	75	— 3	225	675
7	876	— 2	1752	3504
8	237	— 1	237	237
			— 2855	
9	2674	0		
10	527	1	527	527
11	2947	2	5894	11788
12	223	3	669	2007
13	753	4	3012	12048
14	26	5	130	650
15	59	6	354	2124
16	2	7	14	98
	8554		+ 10600	36343

$$U = 9; \quad \eta = \frac{10600 - 2855}{8554} = 0,905; \quad M = 9,905.$$

$$m^2 = \frac{36343}{8554} = 4,2487$$

$$\mu^2 = 4,2487 - 0,8190 = 3,4297$$

$$\mu = 1,85.$$

60. Wir wenden uns jetzt der Besprechung des Falles zu, daß das Kollektiv stetig und in Klassen eingeteilt ist. Die Rechnung erfährt dadurch keine Änderung, aber die Deutung wird eine andere.

Dadurch, daß nun die Klassenmitte als Vertreter aller Argumentwerte der betreffenden Klasse genommen wird, geht man eigentlich von einem stetigen Kollektiv zu einem unstetigen über, und das bewirkt, daß die Rechnung zu einer Näherungsrechnung wird gegenüber derjenigen, die mit den wirklichen Argumentwerten, also mit der primären Verteilungstafel arbeitet. Die Genauigkeit der Näherung wächst so, wie die Klassengröße abnimmt; dem aber ist durch den Umfang des Kollektivs eine Grenze gesetzt.

Für U wird eine Klassenmitte gewählt und die Abweichungen ε von U werden in Klassenintervalle gerechnet, sind also aufeinanderfolgende ganze Zahlen. Infolgedessen ergeben sich auch η, m, μ in Klassenintervallen und müssen zum Schlusse in die ursprünglichen Maße umgesetzt werden.

Was die Wahl von U betrifft, so verlegt man es bei langen Reihen mit Vorteil in jene Klasse, in der man das arithmetische Mittel vermutet; bei kurzen Reihen kann U auch in das eine oder andere Ende der Tafel verlegt werden. Beide Fälle sind in den nachfolgenden Beispielen zur Anschauung gebracht.

Beispiel 1). Aus der Vermessung von 25878 Rekruten der Armee der Vereinigten Staaten hat sich für deren Körperhöhen die nachstehende Verteilungstafel ergeben[1]). Aus ihr sollen M und μ bestimmt werden.

[1]) K. Pearson, Contributions to the Mathematical Theory of Evolution, Phil. Trans. Roy. Soc., A, vol. 186 (1895), p. 385. Vgl. P. Riebesell, Biometrik und Variationsstatistik. Abderhaldens Handbuch der biologischen Arbeitsmethoden, Abt. V, Teil 2, 1. Hälfte, S. 786; P. Riebesell, Mathematische Statistik und Biometrik. Frankfurt a. M. und Berlin 1932, S. 33; K. Daeves, Praktische Großzahl-Forschung, Berlin 1933, S. 38; W. Winkler, Grundriß der Statistik, Bd. I, Berlin 1931, S. 87. Da sich somit dieses Beispiel, das bereits A. Quetelet in seinem Buche Anthropométrie (Brüssel 1870, S. 259) behandelt, in neueren statistischen Büchern vorfindet, ist es beibehalten worden. Ein weiterer Grund hierfür war der, daß sich dieses Beispiel gut zur Demonstrierung der anzustellenden Berechnungen eignet.

In bezug auf militärstatistische Körpergrößenmessungen vgl. auch G. Th. Fechner (Kollektivmaßlehre, Leipzig 1897, S. 388 und 397): Körpergröße von sächsischen und belgischen Rekruten; O. v. Schjerning (Sanitätsstatistische Betrachtungen über Volk und Heer, Berlin 1910, S. 32): Prozentuale Gliederungszahlen von Militärpflichtigen in Preußen; H. Schwiening (Militärsanitätsstatistik, Berlin 1913, S. 179): Untersuchungen über die geographische Verteilung von Soldaten mit einer Körpergröße von 170 cm und mehr in Deutschland; H. Westergaard und H. C. Nybølle (Grundzüge der Theorie der Statistik, Jena 1928, S. 268): Körpergröße italienischer Rekruten; W. Winkler (Grundriß der Statistik, Bd. I, Berlin 1931, S. 21): Körpergröße von Rekruten im Bezirk Mistelbach; vgl. hierzu auch H. Müller (Die Musterung der Wehrmacht, ein Einblick in die Volksgesundheit Der Öffentliche Gesundheitsdienst 1936, Heft 11, S. 409). Ferner sei in diesem Zusammenhang noch hingewiesen auf E. Fischer (Die Rehobother Bastards, Jena 1913): Körpergröße männlicher erwachsener Bastarde aus Rehoboth; vgl. hierzu E. Weber (Einführung in die Variations- und Erblichkeitsstatistik, München 1935, S. 12, 17, 19, 20) und F J. Linders (Geografiska Annaler 1930, Heft 1): Körpergröße schwedischer Männer 1922—1924; vgl. hierzu F. Ringleb (Mathematische Methoden der Biologie. Leipzig 1937, S. 36).

Tab. 40. Körperhöhen amerikanischer Rekruten. Arithmetisches Mittel und mittlere Abweichung.

X in Zoll[1])	z	ε	$z\varepsilon$	$z\varepsilon^2$
51—52	1	— 15	15	225
52—53	1	— 14	14	196
53—54	2	— 13	26	338
54—55	1	— 12	12	144
55—56	3	— 11	33	363
56—57	7	— 10	70	700
57—58	6	— 9	54	486
58—59	10	— 8	80	640
59—60	15	— 7	105	735
60—61	50	— 6	300	1800
61—62	526	— 5	2630	13150
62—63	1237	— 4	4948	19792
63—64	1947	— 3	5841	17523
64—65	3019	— 2	6038	12076
65—66	3475	— 1	3475	3475
			— 23641	
66—67	4054	0		
67—68	3631	1	3631	3631
68—69	3133	2	6266	12532
69—70	2075	3	6225	18675
70—71	1485	4	5940	23760
71—72	680	5	3400	17000
72—73	343	6	2058	12348
73—74	118	7	826	5782
74—75	42	8	336	2688
75—76	9	9	81	729
76—77	6	10	60	600
77—78	2	11	22	242
	25878		+ 28845	169630

$$U = 66{,}5;\quad \gamma = \frac{28845 - 23641}{25878} = 0{,}2011;\quad M = 66{,}5 + 0{,}2011 = 66{,}7011 \text{ Zoll.}$$

$$m^2 = \frac{169630}{25878} = 6{,}5550,$$

$$\mu^2 = 6{,}5550 - 0{,}0404 = 6{,}5146.$$

$$\mu = 2{,}5524 \text{ Zoll.}$$

Mit ausreichender Schärfe kann man das Ergebnis mit

$$M = 66{,}70'',\quad \mu = 2{,}55''$$

ansetzen.

[1]) 1 amerikanischer = 1 englischer Zoll = 25,4 mm.

Beispiel 2). Dieses betrifft eine wirtschaftliche Erfahrungsreihe, nämlich die Sommerarbeitslöhne landwirtschaftlicher Arbeiter in 1200 Gebieten des Deutschen Reiches [1]).

Tab. 41.

X in Mark	Zahl der Gebiete z	ε	$z\varepsilon$	$z\varepsilon^2$
0,60—0,80	7	0	0	0
0,80—1,00	30	1	30	30
1,00—1,20	45	2	90	180
1,20—1,40	104	3	312	936
1,40—1,60	239	4	956	3824
1,60—1,80	247	5	1235	6175
1,80—2,00	257	6	1542	9252
2,00—2,20	50	7	350	2450
2,20—2,40	79	8	632	5056
2,40—2,60	84	9	756	6804
2,60—2,80	19	10	190	1900
2,80—3,00	34	11	374	4114
3,00—3,20	0	12	0	0
3,20—3,40	3	13	39	507
3,40—3,60	1	14	14	196
3,60—3,80	1	15	15	225
	1200		6535	41649

$$U = 0{,}70; \quad \eta = \frac{6535}{1200} = 5{,}4458 \text{ in Klassen} = 1{,}0892 \text{ in M.};$$

$$M = 0{,}70 + 1{,}09 = 1{,}79 \text{ M.}$$

$$m^2 = \frac{41649}{1200} = 34{,}7075,$$

$$\mu^2 = 34{,}7075 - 29{,}6567 = 5{,}0508,$$

$$\mu = 2{,}2474 \text{ in Klassen} = 0{,}45 \text{ M.}$$

61. Das in Art. 38 zur Bestimmung des arithmetischen Mittels entwickelte Summenverfahren kann so fortgebildet werden, daß es auch zur Berechnung der mittleren Abweichung führt; es bietet dem vorstehenden Verfahren gegenüber den Vorteil, daß es Multiplikationen erspart und, bis auf die Schlußrechnung, nur Additionen erfordert. Anknüpfend an die dort gebildeten Summen und Hauptsummen, die in die folgende Tabelle eingetragen sind, setzen wir die Summenbildung, und zwar gleich zweiteilig, fort.

---

[1]) A. Mitscherlich, Die Schwankungen der landwirtschaftlichen Reinerträge. Zeitschr. f. d. ges. Staatswissensch., Ergänzungsheft VIII, Tübingen 1903, S. 11 u. f.

Tabelle der ersten Summen.

x	z	s
$a+1$	z_1	s_1
$a+2$	z_2	s_2
$a+3$	z_3	s_3
.	.	.
.	.	.
.	.	.
.	.	.
$a+k-2$	z_{k-2}	s_{k-2}
$a+k-1$	z_{k-1}	S_1^-
$U=a+k$	z_k	
$a+k+1$	z_{k+1}	S_1^+
$a+k+2$	z_{k+2}	s_{k+2}
.	.	.
.	.	.
.	.	.
.	.	.
$a+n-2$	z_{n-2}	s_{n-2}
$a+n-1$	z_{n-1}	s_{n-1}
$a+n$	z_n	s_n
	$N=\Sigma(z)$	

Von oben

$$
\begin{aligned}
s_1' &= s_1 &&= z_1 \\
s_2' &= s_1+s_2 &&= 2z_1+z_2 \\
s_3' &= s_1+s_2+s_3 &&= 3z_1+2z_2+z_3 \\
&\cdots\cdots\cdots\cdots \\
s_{k-3}' &= s_1+s_2+s_3+\ldots+s_{k-3} &&= (k-3)z_1+(k-4)z_2+ \\
& && \quad +(k-5)z_3+\ldots+z_{k-3}.
\end{aligned}
\tag{5}
$$

Hieraus ergibt sich durch Summierung

$$
s_1'+s_2'+s_3'+\ldots+s_{k-3}'=
$$

$$
=\frac{(k-3)(k-2)}{2}z_1+\frac{(k-4)(k-3)}{2}z_2+\frac{(k-5)(k-4)}{2}z_3+\ldots+z_{k-3}=S_2^-.
$$

nun ist

$$
\begin{aligned}
(k-3)(k-2) &= (k-1-2)(k-1-1)=(k-1)^2-3(k-1)+2 \\
(k-4)(k-3) &= (k-2-2)(k-2-1)=(k-2)^2-3(k-2)+2, \\
&\cdots\cdots\cdots\cdots\cdots\cdots
\end{aligned}
$$

folglich ist

$$
2S_2^- = \sum_1^{k-3}(z\,s^2)+3\sum_1^{k-3}(z\,s)+2\sum_1^{k-3}(z). \tag{6}
$$

Von unten:

$$\left.\begin{aligned}
s'_n &= s_n && = z_n \\
s'_{n-1} &= s_n + s_{n-1} && = 2z_n + z_{n-1} \\
s'_{n-2} &= s_n + s_{n-1} + s_{n-2} && = 3z_n + 2z_{n-1} + z_{n-2} \\
s'_{k+3} &= s_n + s_{n-1} + s_{n-2} + && = (n-k-2)z_n + (n-k-3)z_{n-1} + \\
&\quad + \ldots + s_{k+3} = && \quad + (n-k-4)z_{n-2} + \ldots + z_{k+3}.
\end{aligned}\right\} \quad (7)$$

Daraus erhält man durch Summierung

$$s'_n + s'_{n-1} + s'_{n-2} + \ldots + s'_{k+3} = \frac{(n-k-2)(n-k-1)}{2} z_n + \frac{(n-k-3)(n-k-2)}{2} z_{n-1} + \\ + \frac{(n-k-4)(n-k-3)}{2} z_{n-2} + \ldots + z_{k+3} = S_2^+;$$

die Entwicklung der Zähler führt weiter auf

$$(n-k-2)(n-k-1) = (n-k)^2 - 3(n-k) + 2$$
$$(n-k-3)(n-k-2) = (n-k-1)^2 - 3(n-k-1) + 2$$

infolgedessen wird

$$2S_2^+ = \sum_{k+3}^{n} (z\,\varepsilon^2) - 3\sum_{k+3}^{n} (z\,\varepsilon) + 2\sum_{k+3}^{n} (z). \quad (8)$$

Die Berechnung von m^2 erfordert die Bildung von

$$\sum_1^n (z\,\varepsilon^2);$$

dies aber läßt folgende Auflösung zu:

$$\sum_1^n (z\,\varepsilon^2) = \sum_1^{k-3} (z\,\varepsilon^2) + \sum_{k+3}^{n} (z\,\varepsilon^2) + 4z_{k-2} + z_{k-1} + z_{k+1} + 4z_{k+2};$$

ersetzt man die Summen rechter Hand durch ihre Ausdrücke aus (6) und (8), so wird

$$\left.\begin{aligned}
\sum_1^n (z\,\varepsilon^2) &= 2\left(S_2^+ + S_2^-\right) + \\
&+ 3\left(-\sum_1^{k-3} (z\,\varepsilon) + 2z_{k-2} + \sum_{k+3}^{n} (z\,\varepsilon) + 2z_{k+2}\right) - \\
&- 2\left(\sum_1^{k-3} (z) + z_{k-2} + \sum_{k+3}^{n} (z) + z_{k+2}\right) + z_{k-1} + z_{k+1} \\
&= 2\left(S_2^+ + S_2^-\right) + 3\left(-\sum_1^{k-2} (z\,\varepsilon) + \sum_{k+2}^{n} (z\,\varepsilon)\right) - 2(s_{k-2} + s_{k+2}) + z_{k-1} + z_{k+1} \\
&= 2\left(S_2^+ + S_2^-\right) + 3\left(S_1^+ + S_1^-\right) + \left(S_0^+ + S_0^-\right) = 2\sum\nolimits_2 + 3\sum\nolimits_1 + \sum\nolimits_0,
\end{aligned}\right\} \quad (9)$$

wenn man abkürzend setzt

$$S_\nu^+ + S_\nu^- = \sum\nolimits_\nu. \quad (10)$$

Die Tabelle erhält also, wenn man die Berechnung der mittleren Abweichung einbezieht, die folgende Gestalt:

Schema der Tabelle zur Berechnung des arithmetischen Mittels und der mittleren Abweichung.

x	z	s	s'
$a+1$	z_1	s_1	s'_1
$a+2$	z_2	s_2	s'_2
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
$a+k-3$	z_{k-3}	s_{k-3}	s'_{k-3}
$a+k-2$	z_{k-2}	s_{k-2}	S_2^-
$a+k-1$	z_{k-1}	$S_0^- = s_{k-1},\ S_1^-$	
$U = a+k$	z_k	s_k	
$a+k+1$	z_{k+1}	$S_0^+ = s_{k+1},\ S_1^+$	
$a+k+2$	z_{k+2}	s_{k+2}	S_2^+
$a+k+3$	z_{k+3}	s_{k+3}	s'_{k+3}
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
$a+n-2$	z_{n-2}	s_{n-2}	s'_{n-2}
$a+n-1$	z_{n-1}	s_{n-1}	s'_{n-1}
$a+n$	z_n	s_n	s'_n
	$N = \Sigma(z)$		

Über die Wahl von U ist das Nötige bereits gesagt worden. Verlegt man es auf den Anfang, so entfallen S_0^-, S_1^-, S_2^-, und S_0^+, S_1^+, S_2^+ sind dann auch schon Σ_0, Σ_1, Σ_2; verlegt man es an das Ende, so bleiben nur S_0^-, S_1^-, S_2^- bestehen und bedeuten zugleich Σ_0, Σ_1, Σ_2.

Eine Probe, außer der schon bekannten, ergibt sich aus den Ansätzen

$$\begin{aligned} s'_{k-3} &= (k-3)\ z_1 + (k-4)\ z_2 + (k-5)\ z_3 + \ldots + z_{k-3} \\ s_{k-2} &= \qquad\ z_1 + \qquad\ z_2 + \qquad\ z_3 + \ldots + z_{k-3} + z_{k-2}, \end{aligned}$$

durch Summierung entsteht

$$s'_{k-3} + s_{k-2} = (k-2)\ z_1 + (k-3)\ z_2 + (k-4)\ z_3 + \ldots + z_{k-2},$$

was so viel heißt als

$$s'_{k-3} + s_{k-2} = S_1^-; \tag{11}$$

ebenso hat man im untern Teil der Tafel

$$s'_{k+3} + s_{k+2} = S_1^+. \tag{12}$$

62. Beispiele. 1) Die in Art. 60, 2) vorgeführten Sommerarbeitslöhne werden hier nochmals aufgenommen und nach dem Summenverfahren erledigt, wobei dies einemal neben die Zahlen, die in der Schlußrechnung Verwendung finden, noch die Buchstaben der allgemeinen Entwicklung hingeschrieben sind.

Bestimmung von M und μ nach dem Summenverfahren.

x in Mark	z	s	s'
0,70	7	7	7
0,90	30	37	44
1,10	45	82	126
1,30	104	186	312
1,50	239	425	489 (S_2^-)
1,70	247	(S_0^-) 672, 737 (S_1^-)	
1,90	257	929	
2,10	50	(S_0^+) 271, 473 (S_1^+)	
2,30	79	221	439 (S_2^+)
2,50	84	142	252
2,70	19	58	110
2,90	34	39	52
3,10	.	5	13
3,30	3	5	8
3,50	1	2	3
3,70	1	1	1
	1200		

Proben:

$$929 + 271 = 1200; \quad 425 + 312 = 737; \quad 221 + 252 = 473.$$

$$U = 1,90; \quad \Delta_0 = 271 - 672 = -401, \quad \Delta_1 = 473 - 737 = -264$$

$$\eta = -\frac{401 + 264}{1200} = -0,554 \text{ Klassen} = -0,1108 \text{ M.}, \quad M = 1,90 - 0,1108 = 1,79 \text{ M.}$$

$$\Sigma_0 = 271 + 672 = 943, \quad \Sigma_1 = 473 + 737 = 1210, \quad \Sigma_2 = 439 + 489 = 928$$

$$2\,\Sigma_2 + 3\,\Sigma_1 + \Sigma_0 = 6429$$

$$m^2 \; \frac{6429}{1200} = 5,3575$$

$$\mu^2 = 5,3575 - 0,3069 = 5,0506,$$

$$\mu = 2,2474 \text{ Klassen} = 0,45 \text{ M.}$$

Das vorstehende Zahlenbeispiel von A. Mitscherlich behandelt auch A. Timpe[1]). Er zeigt, daß 70% des Kollektivs in dem Bereich $M - \mu$ bis $M + \mu$ liegen, was auf eine normale Häufigkeitsverteilung hinweist. Weiter ist zu bemerken, daß P. Lorenz[2]) das Zahlenbeispiel von Mitscherlich verwendet, um daran die von ihm aufgefundene Summenmethode zu illustrieren.

2) Bei diesem Beispiel, das sich auf die in Art. 60 zum erstenmal erwähnten amerikanischen Rekruten bezieht, ist die von allem Beiwerk entblößte Rechnung wiedergegeben.

Mittlere Abweichung der Körperhöhen amerikanischer Rekruten.

x in Zoll	z	s		s'
51,5	1	1		1
52,5	1	2		3
53,5	2	4		7
54,5	1	5		12
55,5	3	8		20
56,5	7	15		35
57,5	6	21		56
58,5	10	31		87
59,5	15	46		133
60,5	50	96		229
61,5	526	622		851
62,5	1237	1859		2710
63,5	1947	3806		6516
64,5	3019	6825		10660
65,5	3475	10300,	13341	
66,5	4054	14354		
67,5	3631	11524,	17321	
68,5	3133	7893		17250
69,5	2075	4760		9428
70,5	1485	2685		4668
71,5	680	1200		1983
72,5	343	520		783
73,5	118	177		263
74,5	42	59		86
75,5	9	17		27
76,5	6	8		10
77,5	2	2		2
	25878			

[1]) A. Timpe, Einführung in die Finanz- und Wirtschaftsmathematik. Berlin 1934, S. 169.

[2]) P. Lorenz, Zur Summenmethode. Deutsches Statistisches Zentralblatt 1932. Heft 3. Spalte 69 u. f.

Proben:

$$14354 + 11524 = 25878; \quad 6825 + 6516 = 13341; \quad 7893 + 9428 = 17321.$$

$$U = 66,5; \quad \Delta_0 = 11524 - 10300 = 1224, \quad \Delta_1 = 17321 - 13341 = 3980$$

$$\eta = \frac{5204}{25878} = 0,2011, \quad M = 66,5 + 0,2011 = 66,7011 \text{ Zoll.}$$

$$\Sigma_0 = 21824, \qquad \Sigma_1 = 30662, \qquad \Sigma_2 = 27910$$

$$2\Sigma_2 + 3\Sigma_1 + \Sigma_0 = 169630$$

$$m^2 = \frac{169630}{25878} = 6,5550$$

$$\mu^2 = 6,5550 - 0,0404 = 6,5146$$

$$\mu = 2,5524 \text{ Zoll.}$$

3) In einteiliger Rechnung ist das folgende wesentlich engere Erfahrungsmaterial bearbeitet. Es stellt die Verteilung der mittleren Barometerhöhen in Dresden im Januar des 106jährigen Zeitraums 1828 bis 1933 dar. In dem arithmetischen Mittel wird eine den genannten Monat kennzeichnende meteorologische Zahl erhalten, wie sie für jeden Monat abgeleitet werden könnte. Die mittlere Abweichung ist ein Ausdruck für die Beständigkeit des Luftdrucks in dem betreffenden Monat; läge sie für alle Monate vor, so wäre daraus zu ersehen, wie sich die Beständigkeit im Laufe des Jahres ändert.

Tab. 42. Mittlerer Barometerstand in Dresden im Januar 1828 bis 1933[1]).

x in mm	Zahl der Jahre z	s	s'
743,5	1	106	
745,5	4	105 438	
747,5	2	101	892
749,5	15	99	337
751,5	19	84	238
753,5	20	65	154
755,5	22	45	89
757,5	10	23	44
759,5	6	13	21
761,5	6	7	8
763,5	1	1	1
	106		

[1]) Festschrift der Landeswetterwarte, 16. Tagung zu Dresden 1929, S. XIV; Deutsches Meteorologisches Jahrbuch, Jahrgänge 1928—1933.

Probe:

$$101 + 337 = 438$$

$$U = 743{,}5;\quad \Delta_0 = 105,\quad \Delta_1 = 438;\quad \gamma_1 = 5{,}12264 \text{ Klassen} = 10{,}24528 \text{ mm}$$

$$M = 743{,}5 + 10{,}24528 = 753{,}7453$$

$$\Sigma_0 = 105,\qquad \Sigma_1 = 438,\qquad \Sigma_2 = 892$$

$$2\,\Sigma_2 + 3\,\Sigma_1 + \Sigma_0 = 3203$$

$$m^2 = \frac{3203}{106} = 30{,}2170$$

$$\mu^2 = 30{,}2170 - 26{,}24144 = 3{,}97556$$

$$\mu = 1{,}9939 \text{ Klassen} = 3{,}9878 \text{ mm.}$$

63. Die mittlere Abweichung wird bei einem stetigen, in Klassen eingeteilten Kollektiv so gerechnet, als ob die Werte der Variablen innerhalb einer Klasse gleich und gleich der Klassenmitte wären. Daraus entspringt notwendig ein Fehler in der Berechnung; eine Schätzung seiner Größe und daher eine Berichtigung kann wieder nur durch eine Annahme geschehen, von der man erwarten darf, daß sie sich der Wirklichkeit besser anpaßt als die obige. Sie wird so getroffen, daß sich die Werte von X innerhalb einer Klasse gleichförmig über diese verteilen, daß also die Häufigkeitslinie durch das in Art. 27 (Fig. 2) besprochene Treppenpolygon ersetzt wird.

Sei X wie immer das allgemeine Zeichen für die Variable, x ihr zur Klassenmitte gehöriger Wert, $X - x = \varepsilon$, z die Klassenhäufigkeit; dann gehe man von der Identität aus

$$M - x = M - X + \varepsilon,$$

aus der sich

$$(M - x)^2 = (M - X)^2 + 2\,(M - X)\,\varepsilon + \varepsilon^2$$

ergibt; über das Klassenintervall summiert, liefert dies

$$z\,(M - x)^2 = \Sigma\,(M - X)^2 - 2\,\Sigma\,(X\,\varepsilon) + \Sigma\,(\varepsilon^2),$$

weil $2\,\Sigma\,(M\,\varepsilon) = 2\,M\,\Sigma\,(\varepsilon) = 0$ ist; aber auch die Doppelsumme $\Sigma\,(X\,\varepsilon)$ verschwindet bei gleichförmiger Verteilung der X-Werte in der Klasse. Somit ist

$$\begin{aligned} z\,(M - x)^2 &= \Sigma\,(M - X)^2 + \Sigma\,(\varepsilon^2) \\ &= \Sigma\,(M - X)^2 + z\,\mu_\varepsilon^2; \end{aligned}$$

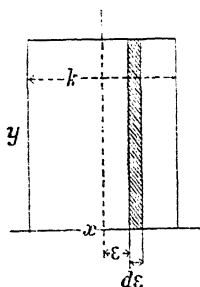
wird dieser Ansatz für alle Klassen gemacht und dann die Summe gezogen, so erhält man

$$\Sigma\,z\,(M - x)^2 = \Sigma\,(M - X)^2 + \mu_\varepsilon^2\,\Sigma\,(z),$$

wobei sich die erste rechtsstehende Summe nunmehr auf das ganze Wertgebiet von X erstreckt. Durch Division mit $\Sigma\,(z)$ erhält man weiter

$$\frac{\Sigma\,z\,(M - x)^2}{\Sigma\,(z)} = \frac{\Sigma\,(M - X)^2}{\Sigma\,(z)} + \mu_\varepsilon^2;$$

das erste Glied ist der nach der üblichen Methode berechnete Wert des Quadrats der mittleren Abweichung, er heiße μ^2; das zweite Glied bedeutet den aus der neuen Annahme hervorgehenden Wert derselben Größe, er heiße μ_1^2; man hat also

$$\mu^2 = \mu_1^2 + \mu_\varepsilon^2.$$

An Hand der Fig. 20 ergibt sich

$$\mu_\varepsilon^2 = \frac{\int_{-\frac{k}{2}}^{\frac{k}{2}} \varepsilon^2 y \, d\varepsilon}{\int_{-\frac{k}{2}}^{\frac{k}{2}} y \, d\varepsilon} = \frac{\int_{-\frac{k}{2}}^{\frac{k}{2}} \varepsilon^2 \, d\varepsilon}{\int_{-\frac{k}{2}}^{\frac{k}{2}} d\varepsilon} = \frac{k^2}{12},$$

Fig. 20. Zur Ableitung der Sheppardschen Formel.

wenn k die Klassengröße bezeichnet. Es ist also schließlich

$$\mu_1^2 = \mu^2 - \frac{k^2}{12}. \tag{13}$$

Die Korrektur beträgt demnach $^1/_{12}$ des Quadrats der Klassengröße, $^1/_{12}$ also, wenn man die Rechnung in Klassengrößen führt. Der Einfluß der ursprünglichen Vorstellung auf das Resultat ist hiernach in der Regel ein sehr geringer. Diese verschärfte Berechnungsweise ist von W. F. Sheppard angegeben worden.[1])

In dem Beispiel, betreffend die Körperhöhen amerikanischer Rekruten (Art. 60, 1) ist bei Klassenrechnung

$$\mu^2 = 6{,}5146$$

gefunden worden; bringt man davon $^1/_{12} = 0{,}5429\ldots$ in Abzug, so ergibt sich

$$\mu_1^2 = 5{,}9717$$

und $\mu_1 = 2{,}4437$ gegenüber $\mu = 2{,}5524$.

Bei großen Klassenintervallen wird man die Sheppardsche Korrektur zu berücksichtigen haben.

64. Die durchschnittliche Abweichung. Unter dieser Bezeichnung wird als ein zweites gebräuchliches Streuungsmaß das arithmetische Mittel der absoluten Werte der Abweichungen von einem Mittelwerte verstanden. Als solcher kann das arithmetische Mittel, aber auch der Zentralwert benützt werden. Die letztere Wahl hätte aus einem gleich zu erwähnenden theoretischen Grunde den Vorzug vor der ersteren, doch wird auch hier zumeist an dem arithmetischen Mittel festgehalten.

[1]) Journ. Roy. Statist. Soc., vol. IX (1897), p. 698.

Das arithmetische Mittel wurde als derjenige Vergleichswert erkannt, der zu der kleinsten mittleren Abweichung führt. In gleicher Weise kommt dem Zentralwert die besondere Eigenschaft zu, die kleinste durchschnittliche Abweichung zu geben. Dies geht aus folgender Überlegung hervor. Der Ausgangswert U werde so gewählt, daß m von den beobachteten Argumentwerten über ihm, $n-m$ unter ihm liegen; bei dieser Lage sei Θ die durchschnittliche Abweichung. Verschiebt man U nach rechts um eine so kleine Strecke u, daß die Verteilung der Argumentwerte dadurch nicht verändert wird, so hat sich die obere Summe der Abweichungen um mu vermindert, die untere aber um $(n-m)u$ vermehrt, infolgedessen ist die durchschnittliche Abweichung bei dem neuen Ausgangswert

$$\Theta + \frac{(n-m)u - mu}{n} = \Theta + \frac{u}{n}(n-2m);$$

sie wäre

$$\Theta + \frac{mu-(n-m)u}{n} = \Theta + \frac{u}{n}(2m-n),$$

wenn die Verschiebung unter dem gleichen Vorbehalt nach links erfolgte; das erstemal erfährt Θ eine Vergrößerung, wenn $m < \frac{n}{2}$, das zweitemal, wenn $m > \frac{n}{2}$; beidemal tritt die Vergrößerung nicht ein und Θ stellt ein Minimum dar, wenn $m = \frac{n}{2}$, d. h. wenn U so liegt, daß je die Hälfte der Argumentwerte unter ihm und über ihm liegt; diese Eigenschaft aber kennzeichnet den Zentralwert.

Wir geben der durchschnittlichen Abweichung das Zeichen ϑ, wozu noch beigefügt werden muß, ob sie sich auf C oder M bezieht; wir denken, wenn nichts bemerkt wird, an M. Ihre Berechnung aus der Verteilungstafel vollzieht sich in folgender Weise. Erklärt soll dies werden an der Tafel der Körperhöhen amerikanischer Rekruten. Hat man die Rechnung nach dem Schema Art. 60, 1) angelegt, so ist $U = 66{,}5$, die Summe der Abweichungen von diesem Ausgangswerte im oberen Teil der Tafel 28845, die Summe der absoluten Werte der Abweichungen im unteren Teil der Tafel 23641, folglich die Gesamtsumme

$$52486;$$

es handelt sich nun darum, um welchen Betrag sich diese Summe ändert, wenn man statt U nimmt $M = 66{,}7011$, das um 0,2011 größer ist; über dem Ausgangswert liegen laut der Tafel (S. 137) 11524, unter ihm 10300 Glieder, also ändert sich die Summe um

$$(10300 + 4054 - 11524) \cdot 0{,}2011,$$

d. i. um 569,11, wird also

$$52486 + 569{,}11 = 53055{,}11;$$

somit ist

$$\vartheta = \frac{53055{,}11}{25878} = 2{,}0502''.$$

Geht man von C aus, das mit 66,6509 gefunden wird und somit nur 0,1509 größer ist als U, so bleibt die Sachlage die gleiche wie vorhin und die Änderung der Abweichungssumme beträgt

$$(10300 + 4054 - 11524) \cdot 0{,}1509,$$

d. i. 427,05, die Summe wird

$$52486 + 427{,}05 = 52913{,}05$$

und

$$\vartheta = \frac{52913{,}05}{25878} = 2{,}0447'',$$

tatsächlich kleiner als bei Zugrundelegung von M.

Arbeitet man nach dem Summenverfahren, so gestaltet sich die Rechnung wie folgt. Es ist

$$\Sigma z|\varepsilon| = \sum_1^{k-2} z|\varepsilon| + z_{k-1} + z_{k+1} + \sum_n^{k+2}(z\varepsilon);$$

den Gleichungen (8) bis (11) des Art. 38 zufolge ist aber

$$\sum_1^{k-2} z|\varepsilon| = -\sum^{k-2} z\varepsilon = S_1^- + s_{k-2}$$

$$\sum_n^{k+2}(z\varepsilon) = \qquad S_1^+ + s_{k+2},$$

folglich weiter

$$\begin{aligned}\Sigma z|\varepsilon| &= S_1^- + s_{k-2} + z_{k-1} + z_{k+1} + s_{k+2} + S_1^+ \\ &= S_1^- + S_0^- + S_0^+ + S_1^+ \\ &= \Sigma_0 + \Sigma_1,\end{aligned}$$

wenn man sich wieder der Abkürzung (10) in Art. 61 bedient. Schließlich ist also

$$\frac{\Sigma_0 + \Sigma_1}{n}$$

die durchschnittliche Abweichung von U, die nun noch zu korrigieren ist wegen des Unterschiedes zwischen U und M oder U und C, wofür die Tabelle die erforderlichen Daten enthält.

Dasselbe Beobachtungsmaterial wie im vorigen Beispiel ist im Art. 62, 2 nach dem Summenverfahren behandelt worden; es fand sich dort

$$\Sigma_0 = 21824, \qquad \Sigma_1 = 30662;$$

somit ist $\Sigma_0 + \Sigma_1 = 52486$ in Übereinstimmung mit dem vorhin gefundenen Werte. Alles übrige verläuft wie oben mit Hilfe der Zahlen 10300, 4054 und 11524, die die Tafel auch angibt.

65. Quartile und Perzentile. Ein weiteres Mittel, die Streuung zu kennzeichnen, bilden die Quartile, die in früherer Zeit besonders in der Anthropologie verwendet wurden. Ihre Definition ist die folgende.

Man bezeichne mit Q_1 jenen Argumentwert, unter welchen ein Viertel der Kollektivglieder fällt, so daß Dreiviertel über ihn hinausgehen. Man nenne ferner Q_3 denjenigen Argumentwert, über welchen ein Viertel der Kollektivglieder fällt, so daß Dreiviertel unter ihm liegen. Das zwischen beiden liegende zweite Quartil ist der Zentralwert C, so daß durch Q_1, C, Q_3 das ganze Kollektiv in vier der Häufigkeit nach gleiche Teile zerfällt.

Bei vollkommener Symmetrie müßte

$$C - Q_1 = Q_3 - C$$

ausfallen; bei linksseitiger Asymmetrie wird

$$C - Q_1 < Q_3 - C,$$

bei rechtsseitiger

$$C - Q_1 > Q_3 - C$$

sein. Da vollkommene Symmetrie zu den Ausnahmen gehört, empfiehlt es sich, als einheitliches Maß die halbe Summe der beiden Abstände $C - Q_1$ und $Q_3 - C$, d. i. $\frac{Q_3 - Q_1}{2}$ mit dem Zeichen Q:

$$Q = \frac{Q_3 - Q_1}{2} \tag{14}$$

einzuführen.

Man nennt Q_1 das erste oder untere, Q_3 das dritte oder obere, Q das Quartil schlechtweg. Man beachte, daß Q sich im allgemeinen auf keinen ausgezeichneten Argumentwert bezieht und nur eine Vorstellung gibt von dem Intervall zwischen unterem und oberem Quartil. Herrscht jedoch vollkommene Symmetrie, so bezieht sich Q auf den Zentralwert, zugleich arithmetisches Mittel, und fällt zusammen mit dem Begriff des wahrscheinlichen Fehlers in der Fehlertheorie.

Die Bestimmung der Quartile geht genau so vor sich wie die Bestimmung von C; nur ist bei Q_1, das man vom unteren Ende aus rechnen wird, von $\frac{N}{4}$ auszugehen, bei Q_3 von $\frac{3N}{4}$ oder $\frac{N}{4}$, je nachdem man es vom unteren oder oberen Ende rechnen will. Dabei tritt wieder der Vorteil des Summenverfahrens hervor.

Bei unstetigen Kollektiven gilt von den Quartilen dasselbe, was von dem Zentralwert gesagt worden ist (Art. 41).

Der Quartilbegriff läßt sich verallgemeinern; man kann statt der Vierteilung auch jede andere Teilung vornehmen; bei Zehnteilung kommt man zu neun Dezilen, deren fünftes mit C zusammenfällt, und bei Hundertteilung zu 99 Perzentilen, deren fünfzigstes wieder C ist. Man hat darin nichts anderes als eine besondere Art der Klasseneinteilung zu erblicken, die sich wegen der Einheitlichkeit zu vergleichenden Betrachtungen über die Streuung eignet.

Als erstes Illustrationsbeispiel wählen wir wieder die Tafel der Körperhöhen amerikanischer Rekruten in der Anlage, die ihr in Art. 62, 2 gegeben worden ist.

Es ist $N = 25878$, $\frac{N}{4} = 6469{,}5$; die nächstkleinere Summe von oben ist 3806, reichend bis $X = 64$; also ist

$$Q_1 = 64 + \frac{6469{,}5 - 3806}{3019} = 64{,}882'';$$

die nächstkleinere Summe von unten, reichend bis $X = 69$, beträgt 4760; also ist

$$Q_3 = 69 - \frac{6469{,}5 - 4760}{3133} = 68{,}454''.$$

Da sich $C = 66{,}651''$ ergibt, so besteht zwischen $C - Q_1 = 1{,}769$ und $Q_3 - C = 1{,}803$ eine Differenz im Betrage von $0{,}034''$, entsprechend der schwachen linksseitigen Asymmetrie. Schließlich ist

$$Q = \frac{68{,}454 - 64{,}882}{2} = 1{,}786''.$$

Bei dem nun vorzuführenden zweiten Beispiel muß auf die Klassengröße Rücksicht genommen werden; Gegenstand desselben sind die in den Art. 60, 2 und 62, 1 angeführten Sommerarbeitslöhne landwirtschaftlicher Arbeiter.

Hier ist

$$N = 1200, \quad \frac{N}{4} = 300,$$

$$Q_1 = 1{,}40 + \frac{300 - 186}{239} \cdot 0{,}2 = 1{,}50 \text{ M.}$$

$$Q_3 = 2{,}00 - \frac{300 - 271}{257} \cdot 0{,}2 = 1{,}98 \text{ M.}$$

$$C = 1{,}74, \qquad C - Q_1 = 0{,}24, \qquad Q_3 - C = 0{,}24$$

$$Q = \frac{1{,}98 - 1{,}50}{2} = 0{,}24 \text{ M.}$$

Die Unregelmäßigkeit dieser Verteilung, die sich in dem raschen Abfall von der Lohnstufe von 1,8 bis 2,0 auf 2,0 bis 2,2 M äußert, verrät sich auch dadurch, daß bei linksseitiger Asymmetrie $C - Q_1 = Q_3 - C$ ist gegen die Regel.

Die Verwendung von Dezilen soll an Sterbetafeln gezeigt werden. Eine Sterbetafel kann als eine Verteilung der Sterbefälle nach Altersklassen aufgefaßt werden; die Zahlen der Lebenden bilden dabei die Summenreihe vom untern Ende aus. Die drei Mittelwerte M, C, D bezeichnen der Reihe nach die mittlere, die wahrscheinliche Lebensdauer und das Lexissche Normalalter[1]).

Nachstehend sind die Dezilen $d_1, d_2, \ldots d_9$ der „Allgemeinen deutschen Sterbetafel für das Jahr 1933"[2]) für beide Geschlechter zusammengestellt.

---

[1]) Vgl. A. Timpe, Zur Lexisschen Theorie der Lebensdauer. Festschrift zu Ehren von G. Höckner. Berlin 1935, S. 68 u. f.

[2]) Sonderheft zu Wirtschaft und Statistik Nr. 15. Berlin 1935, S. 62.

Tab. 43. Dezile der Allgemeinen deutschen Sterbetafel für das Jahr 1933.

	Männlich	Weiblich
d_1	3,79 Jahre	17,66 Jahre
d_2	43,65 „	49,24 „
d_3	56,90 „	60,58 „
d_4	63,85 „	71,06 „
d_5	68,66 „	74,49 „
d_6	72,55 „	80,87 „
d_7	76,00 „	84,77 „
d_8	79,44 „	66,76 „
d_9	83,50 „	77,62 „

Die Tabelle läßt bezeichnende Unterschiede in der Sterblichkeit der beiden Geschlechter erkennen: die günstigere Lage des weiblichen Geschlechtes am Lebensbeginn, insbesondere aber im jugendlichen Alter, seine ständige Bevorzugung vor dem männlichen.

Da es viel Raum erfordern würde, die Sterbetafeln hier zum Abdruck zu bringen, so sei die Rechnung nur an einer Position aus der Tafel des männlichen Geschlechtes erklärt:

Altersklasse	Sterbefälle	Lebende am Beginn der Altersklasse
63—64	1736	61473
64—65	1871	59737

Dem vierten Dezil entsprechen, da 100000 die Basis der Tafel, 60000 Lebende; folglich ist

$$d_4 = 64 - \frac{60000-59737}{1736} = 63{,}85 \text{ Jahre.}$$

66. Sollen mittlere Abweichung, durchschnittliche Abweichung und Quartil gleichberechtigte Maße der Streuung sein, so muß sich die gegenseitige Stellung zweier Verteilungen als gleich ergeben, ob man sie nach dem einen oder dem andern Maße beurteilt. Selbstverständlich wird es sich bei solchen vergleichenden Untersuchungen, und sie sind ein wesentlicher Behelf der Forschung, immer nur um gleiche oder ähnliche Materien handeln, und solche zeigen meist auch ähnliche Züge in der Verteilung. Wäre für die betreffende Materie die Häufigkeitskurve analytisch bestimmt, so ließe sich die Frage nach der Gleichberechtigung der drei Streuungsmaße rein theoretisch erledigen. Ist dem nicht so, so kann dies nur auf empirischem Wege geschehen. Da man naturgemäß die Streuung als dem Streuungsmaß proportional setzen wird, so müßten auch die Streuungsmaße, aus verschiedenen Verteilungen abgeleitet, paarweise konstante Verhältnisse bilden.

Wir wollen hier eine solche empirische Prüfung durchführen.

Nachstehend sind die Verteilungen der Körpergrößen der Knaben und Mädchen im Alter von 9 bis 10 Jahren in der Stadt Bern[1]) mitgeteilt; aus den Zahlenreihen sind die Streuungsmaße und der Zentralwert, um mit ihm die Quartile vergleichen zu können, abgeleitet worden.

Tab. 44. Vergleichung der Streuungsmaße der Körpergröße von Knaben und Mädchen.

Körpergröße in cm X	Knaben z	Mädchen z
117—121	3	8
121—125	21	28
125—129	74	82
129—133	146	140
133—137	212	188
137—141	192	148
141—145	82	69
145—149	28	15
149—153	7	16
	765	694

Es ergaben sich folgende Resultate:

Knaben		Mädchen	
$\mu = 5{,}694$ cm	$Q_3 = 139{,}44$ cm	$\mu = 6{,}191$ cm	$Q_3 = 139{,}00$ cm
$\vartheta = 4{,}47$ „	$Q = 3{,}94$ „	$\vartheta = 4{,}72$ „	$Q = 4{,}20$ „
$Q_1 = 131{,}56$ „	$C - Q_1 = 4{,}04$ „	$Q_1 = 130{,}60$ „	$C - Q_1 = 4{,}28$ „
$C = 135{,}60$ „	$Q_3 - C = 3{,}84$ „	$C = 134{,}88$ „	$Q_3 - C = 4{,}12$ „

Alle drei Maße führen zu dem gleichen Urteil, nämlich, daß die Streuung bei den Mädchen größer ist als bei den Knaben, daß bei ersteren also die Körpergröße stärkeren Variationen unterworfen ist.

Die Verhältnisse der Streuungsmaße sind die folgenden:

Knaben	Mädchen
$\frac{\vartheta}{\mu} = 0{,}78$	$\frac{\vartheta}{\mu} = 0{,}76$
$\frac{Q}{\mu} = 0{,}69$	$\frac{Q}{\mu} = 0{,}68.$

Im vorstehenden Falle handelt es sich um Kollektive von verhältnismäßig kleinem Umfang. Wir stellen ihnen zwei andere, erheblich umfangreichere gegenüber, die Verteilungen der Körperhöhe erwachsener männlicher Personen englischer und schottischer Abkunft.

[1]) K. Müllу, Körperentwicklung von Volksschülern. Zürich 1933, S. 464, 473.

Tab. 45. Vergleichung der Streuungsmaße bei Engländern und Schotten [1]).

X in Zoll	Engländer z	Schotten z
57—58	1	.
58—59	3	1
59—60	12	.
60—61	39	2
61—62	70	2
62—63	128	9
63—64	320	19
64—65	524	47
65—66	740	109
66—67	881	139
67—68	918	210
68—69	886	210
69—70	753	218
70—71	473	115
71—72	254	102
72—73	117	69
73—74	48	26
74—75	16	15
75—76	9	6
76—77	1	4
77—78	1	1
	6194	1304

Aus diesen Daten fließen folgende Ergebnisse:

Engländer		Schotten	
$\mu = 2{,}57$ Zoll	$Q_3 = 69{,}16$ Zoll	$\mu = 2{,}50$ Zoll	$Q_3 = 70{,}10$ Zoll
$\vartheta = 2{,}05$ „	$Q = 1{,}78$ „	$\vartheta = 1{,}95$ „	$Q = 1{,}56$ „
$Q_1 = 65{,}61$ „	$C - Q_1 = 1{,}80$ „	$Q_1 = 66{,}99$ „	$C - Q_1 = 1{,}55$ „
$C = 67{,}41$ „	$Q_3 - C = 1{,}75$ „	$C = 68{,}54$ „	$Q_3 - C = 1{,}56$ „

Alle drei Maße führen zu demselben Urteil und lassen die Körpergröße der Engländer weniger stabil erscheinen als die der Schotten.

Die Verhältnisse stellen sich wie folgt:

Engländer	Schotten
$\frac{\vartheta}{\mu} = 0{,}80$	$\frac{\vartheta}{\mu} = 0{,}78$
$\frac{Q}{\mu} = 0{,}69$	$\frac{Q}{\mu} = 0{,}62.$

[1]) G. U. Yule, An Introduction to the Theory of Statistics. London 1932. S. 88.

Aus beiden Zusammenstellungen geht mit Deutlichkeit die Beziehung $\frac{\vartheta}{\mu} > \frac{Q}{\mu}$ hervor, und die Werte der beiden Verhältnisse gehen in den vier untersuchten Fällen nicht wesentlich auseinander.

67. Mittlere und durchschnittliche Abweichung und Quartil sind absolute Streuungsmaße und darum für Vergleichszwecke nicht immer verwertbar. Kommt es auf die Vergleichung zweier Materien in Bezug auf die Streuung an, der ihre Individuen unterliegen, so muß man bedenken, daß große Maße auch großen Abweichungen, kleine Maße auch entsprechend kleineren Abweichungen ausgesetzt sind, daß sich also dort auch eine größere mittlere Abweichung ergeben wird als hier. Daraus aber darf nicht geschlossen werden, daß auch die Variabilität, die Unbeständigkeit der Maße dort größer ist als hier. Dieser Umstand spielt z. B. eine Rolle, wenn es sich um die Untersuchung verschiedener Rassen von Menschen und Tieren handelt, er kommt ferner in Betracht, wenn vergleichende Untersuchungen über die beiden Geschlechter angestellt werden; die gleichartigen Organe haben hier verschiedene Größe, und darum geben die absoluten Streuungsmaße keine zutreffende Vorstellung von dem Grade der Variabilität[1]).

Diese Erwägungen führen dazu, daß es für vergleichende Untersuchungen notwendig ist, ein relatives Maß der Streuung aufzustellen, von dem gefordert werden muß, daß es auf die Größe der zu vergleichenden Objekte in einer der Natur der Sache angemessenen Weise Rücksicht nehme.

Als ein solches Maß hat Pearson den Variabilitätskoeffizienten in Vorschlag gebracht[2]), der das prozentuale Verhältnis der mittleren Abweichung zum arithmetischen Mittel ausdrückt; bezeichnet man ihn mit V, so ist also

$$V = 100 \frac{\mu}{M}. \qquad (15)$$

Dieser Wahl liegt die Überlegung zugrunde, daß einerseits die Streuung oder Variabilität um so größer zu veranschlagen ist, je größer die mittlere Abweichung, hingegen um so kleiner, je größer die verglichenen Objekte im Mittel sind.

Wir haben im Art. 66 Schulkinder und erwachsene Personen auf die Verteilung ihrer Körperhöhen untersucht. Um die Variabilitätskoeffizienten zu finden, brauchen wir zu den dort angeführten Daten noch die arithmetischen Mittel; diese ergeben sich wie folgt:

Knaben	135,5 cm	Engländer	67,37′
Mädchen	134,9 „	Schotten	68,61′

[1]) Galton hat bei seinen Erblichkeitsforschungen bei Menschen festgestellt, daß die Mittelwerte analoger Maße beim männlichen und weiblichen Geschlecht mit großer Annäherung im Verhältnis 13 : 12 zueinander stehen, so daß 1,08 die Umwandlungszahl weiblicher Maße in die entsprechenden männlichen wäre. (Vgl. F. Galton, Natural Inheritance. London 1889, S. 42.)

[2]) K. Pearson, Mathematical Contributions to the Theory of Evolution. Phil. Trans. Roy. Soc., A, vol. 187 (1896), p. 277.

Nach diesen Mittelwerten sind die Knaben im Durchschnitt etwas größer als die Mädchen. Dasselbe gilt für die Schotten im Vergleich zu den Engländern. Sicheren Aufschluß geben die Variabilitätskoeffizienten, für die sich folgende Werte ergeben:

Knaben	4,20	Engländer	3,81
Mädchen	4,59	Schotten	3,64

Die vorstehenden Variabilitätskoeffizienten geben zugleich auch eine Antwort auf die Frage, ob die Variabilität im Kindesalter eine wesentlich andere ist als bei Erwachsenen.

Man findet durch Vergleichung, daß die Variabilität im Kindesalter erheblich größer ist als bei Erwachsenen.

68. Für die Form der Verteilung eines Kollektivs ist der Grad ihrer Abweichung von der Symmetrie eines der wesentlichsten Merkmale. Vollkommene Symmetrie wird sich praktisch nur ganz ausnahmsweise herausstellen und auch da nur eine Folge abgekürzter Rechnung sein. Während nämlich bei vollkommener Symmetrie die drei Mittelwerte M, C, D zusammenfallen, gehen sie, sobald eine Abweichung von der Symmetrie stattfindet, auseinander; dies kann in so geringem Maße geschehen, daß der Unterschied in den Mittelwerten erst in so späten Dezimalstellen sich bemerkbar macht, daß er im Hinblick auf die vorhandene Schärfe der Messungen gar nicht zur Geltung kommt, so daß man von Symmetrie spricht, wiewohl sie nicht streng vorhanden zu sein braucht.

Bei stärkeren Graden von Asymmetrie tritt eine deutliche Divergenz der Mittelwerte ein, und unter gewissen Voraussetzungen, die in den meisten Fällen erfüllt sind (Art. 42, Schluß), gehen arithmetisches Mittel und dichtester Wert am weitesten auseinander, und man kann in ihrer Differenz ein absolutes Maß der Asymmetrie erblicken.

Dieselben Erwägungen, die hinsichtlich der Streuung angestellt worden sind, führen auch hier zur Festsetzung eines relativen Maßes der Asymmetrie, und als solches wird nach dem Vorschlage Pearsons[1]) die Schiefe der Verteilung verwendet. Es ist dies der Quotient aus dem Unterschied zwischen dem arithmetischen Mittel und dem dichtesten Wert durch die mittlere Abweichung, also der Ausdruck

$$\frac{M-D}{\mu}. \tag{16}$$

Man hat nämlich zu bedenken, daß $M-D$ bei großer Ausbreitung einen erheblichen Betrag erreichen kann, ohne daß die Asymmetrie dabei besonders hervortreten würde, während bei geringer Ausbreitung eine selbst stark erkennbare Asymmetrie von einem sehr kleinen Unterschied der beiden Mittelwerte begleitet sein kann. Es ist also gerechtfertigt, daß man ein Streuungsmaß als Nenner verwendet.

Linksseitige Asymmetrie hat $M > D$ zur Folge, die Schiefe fällt positiv aus, sie wird negativ bei rechtsseitiger Asymmetrie, weil dann $M < D$ ist. Darum

---

[1]) K. Pearson, Contributions to the Mathematical Theory of Evolution. Phil. Trans. Roy. Soc., A, vol. 186 (1895), p. 370.

unterscheidet man die beiden Arten von Asymmetrie auch als positive (links) und negative (rechts). In der Regel ist die Schiefe ein echter Bruch[1]).

Für einige der vorgeführten Verteilungen wollen wir die Schiefe aus den zu ihrer Berechnung erforderlichen Elementen ableiten:

1. Körpergröße amerikanischer Rekruten (Art. 60, 1)):
   $M = 66{,}701''$, $D = 66{,}578''$, $\mu = 2{,}552''$; Schiefe $= +0{,}048$.

2. Körpergröße 9- bis 10jähriger Knaben (Art. 66):
   $M = 135{,}518$ cm, $D = 136{,}068$ cm, $\mu = 5{,}694$ cm; Schiefe $= -0{,}097$.

3. Körpergröße 9- bis 10jähriger Mädchen (Art. 66):
   $M = 134{,}856$ cm, $D = 135{,}180$ cm, $\mu = 6{,}191$ cm; Schiefe $= -0{,}052$.

4. Körpergröße erwachsener Engländer (Art. 66):
   $M = 67{,}373''$, $D = 67{,}536''$, $\mu = 2{,}567''$; Schiefe $= -0{,}063$.

5. Körpergröße erwachsener Schotten (Art. 66):
   $M = 68{,}608''$, $D = 69{,}072''$, $\mu = 2{,}496''$; Schiefe $= -0{,}186$.

6. Alter an Typhoid-Fieber eingelieferter Kranken (Art. 46, 1)):
   $M = 18{,}97$, $D = 14{,}66$, $\mu = 9{,}88$ Jahre; Schiefe $= +0{,}436$.

7. Alter der an Zuckerkrankheit Gestorbenen (Art. 46, 2)):
   $M = 58{,}2$, $D = 64.5$, $\mu = 16{,}1$ Jahre; Schiefe $= -0{,}39$.

69. Zum Schlusse dieses vierten Paragraphen wollen wir zwei vollständig durchgerechnete Beispiele vorführen, in welchen alle bisher zur Sprache gebrachten Größen zur Bestimmung gelangen sollen und die die kürzeste Anordnung der Rechnung nach dem Summenverfahren zeigen.

[1]) Zu einem weiteren Maß der Schiefe gelangt man auf dem Wege der Momentbildung. Bei einer symmetrischen Verteilung sind alle ungeraden Momente, bezogen auf das arithmetische Mittel, gleich Null. Je größer die Schiefe der Verteilung ist, umso größer sind die Werte der ungeraden Momente. Dies führt dazu, die Schiefe der Verteilung durch ungerade Momente zu kennzeichnen. Da das Moment ersten Grades verschwindet, liegt es nahe, das Moment dritten Grades zu verwenden. Dividiert man dieses durch $M_2'^{3/2} = \mu^3$, so erhält man einen Quotienten, der homogen und unabhängig von dem Abszissenmaßstab ist. Wir gelangen so zu folgendem Maß für die

$$\text{Schiefe} = \frac{M_3'}{M_2'^{3/2}} = \frac{M_3'}{\mu^3}.$$

(Vgl. hierzu C. V. L. Charlier, Vorlesungen über die Grundzüge der mathematischen Statistik. Lund 1920, S. 70 u. f.; R. v. Mises, Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik. Leipzig und Wien 1931, S. 240 u. f. und O. Anderson, Einführung in die mathematische Statistik. Wien 1935, S. 161 u. f.)

1) Gegenstand dieses Beispiels sind die täglichen Barometerstände in Cambridge, abgelesen je um 9 Uhr vormittags durch 13 Jahre, im ganzen also $13 \cdot 365 + 3 = 4748$ Ablesungen. Die Klassengröße beträgt 0,1 Zoll[1]).

Tab. 46. Tägliche Barometerstände in Cambridge.[2])

x in Zoll	z	s		s'
28,3	1	1		1
28,4	.	1		2
28,5	.	1		3
28,6	1	2		5
28,7	2	4		9
28,8	6,5	10,5		19,5
28,9	10,5	21		40,5
29,0	23	44		84,5
29,1	24	68		152,5
29,2	63,5	131,5		284
29,3	81	212,5		496,5
29,4	127	339,5		836
29,5	213	552,5		1388,5
29,6	289	841,5		2230
29,7	388	1229,5		3459.5
29,8	479,5	1709		9011.5
29,9	537,5	2246,5	5168,5	
30,0	586	2832,5		
30,1	550	1915,5	3241.5	
30,2	488	1365,5		3614
30,3	350,5	877.5		1876
30,4	246	527		998,5
30,5	150	281		471,5
30,6	85,5	131		190,5
30,7	35	45,5		59,5
30,8	7,5	10,5		14
30,9	2,5	3		3.5
31,0	0,5	0,5		0,5
	4748			

[1]) 1 englischer Zoll = 25,4 mm.

[2]) K. Pearson und A. Lee, On the Distribution of Frequency of the Barometric Height at Divers Stations. Phil. Trans. Roy. Soc., A., vol. 190 (1897), p. 429.

Kontrollen:

1. 2832,5 + 1915,5 = 4748; 1709 + 3459,5 = 5168,5, 1365,5 + 1876 = 3241,5

2.
$$\begin{array}{rlrlrl} & 1915{,}5 & & 3241{,}5 & & 3614 \\ & 2246{,}5 & & 5168{,}5 & & 9011{,}5 \\ \hline \Sigma_0 = & 4162 & \Sigma_1 = & 8410 & \Sigma_2 = & 12625{,}5 \\ \Delta_0 = & -331 & \Delta_1 = & -1927 & & \end{array}$$

3. $\eta = \frac{\Delta_0 + \Delta_1}{N} = -0{,}4756$ Klassen $= -0{,}0476''$

$M = 30 - 0{,}0476 = 29{,}9524''$ (= 760,7910 mm)

4. $\frac{N}{2} = 2374$; $C = 29{,}95 + \frac{2374 - 2246{,}5}{586} \cdot 0{,}1 = 29{,}9718$ (= 761,2837 mm)

5.
$$\begin{array}{rrr} 537{,}5 & 48{,}5 & -84{,}5 \\ 586 & -36 & \\ 550 & & \end{array}$$

$D = 29{,}95 + \frac{48{,}5}{84{,}5} \cdot 0{,}1 = 30{,}0074''$ (= 762,1880 mm)

6. $D - M = 0{,}0550$, $C - M = 0{,}0194$, $3(C - M) = 0{,}0582$

7. $\Sigma_0 + 3\Sigma_1 + 2\Sigma_2 = 54643$, $m^2 = 11{,}5086$

$\mu = 0{,}1\sqrt{11{,}5086 - (0{,}4756)^2} = 0{,}336''$

8. $\Sigma_0 + \Sigma_1 = 12572$; $\vartheta = \frac{12572 + 121{,}38}{4748} \cdot 0{,}1 = 0{,}267''$

9. $\frac{N}{4} = 1187$; $Q_1 = 29{,}65 + \frac{1187 - 841{,}5}{388} \cdot 0{,}1 = 29{,}739''$

$Q_3 = 30{,}25 - \frac{1187 - 877{,}5}{488} \cdot 0{,}1 = 30{,}187''$

$Q = 0{,}224''$

$C - Q_1 = 0{,}233''$ $Q_3 - C = 0{,}215''$

10. $\frac{\vartheta}{\mu} = 0{,}795$ $\frac{Q}{\mu} = 0{,}667$

11. $V = 100\,\frac{0{,}336}{29{,}952} = 1{,}122$

12. Schiefe $= -\frac{0{,}055}{0{,}336} = -0{,}1637$.

Die kritische Würdigung dieser Resultate sei dem Leser überlassen.

Als Hauptergebnisse können die folgenden Werte hingestellt werden.

$M = 29{,}9524''$ $\mu = 0{,}336''$ $V = 1{,}122$
$C = 29{,}9718''$ $\vartheta = 0{,}267''$ Schiefe $= -0{,}1637$.
$D = 30{,}0074''$ $Q = 0{,}224''$

Das Beispiel ist einer großen Untersuchung Pearsons über die Barometerhöhen auf 20 Küstenstationen Englands und 3 weiteren Stationen entnommen; die kürzeste Beobachtungsdauer auf einer Station betrug 5, die längste 13 Jahre.

Die Untersuchung ergab eine Reihe wichtiger Resultate, von welchen angeführt sei die rechtsseitige Asymmetrie, welche der Verteilung der Barometerhöhen eigen ist, die große Beständigkeit insbesondere des dichtesten Wertes (er bewegte sich auf den 23 Stationen zwischen den Grenzen 29,9232 und 30,0493, die nur 0,1261″ voneinander abweichen, während der Zentralwert zwischen den etwas weiteren Grenzen 29,8457 und 29,9834, Differenz 0,1377, variierte). Pearsons Hauptziel war die Anpassung der beobachteten Verteilungen an eine Häufigkeitskurve (vgl. Art. 33, Schluß).

2) Dem bevölkerungsstatistischen Gebiet gehört die Materie an, auf die sich die folgende Tafel bezieht. Sie enthält die Altersverteilung der heiratenden Frauen, die mit Männern im Alter von 25 bis 26 Jahren im Jahre 1931 im Deutschen Reich die Ehe eingegangen sind. Siehe Tab. 47, S. 154.

Kontrollen:

1. 26993 + 21623 = 48616; 13134 + 13272 = 26406; 15087 + 28996 = 44083

2.
$$\begin{array}{rrr} 21623 & 44083 & 90798 \\ 19837 & 26406 & 21906 \\ \hline \Sigma_0 = 41460 & \Sigma_1 = 70489 & \Sigma_2 = 112704 \\ \Delta_0 = 1786 & \Delta_1 = 17677 & \end{array}$$

3. $\eta = \frac{\Delta_0 + \Delta_1}{N} = 0{,}400, \qquad M = 23{,}5 + 0{,}400 = 23{,}900$ Jahre

4. $\frac{N}{2} = 24308; \qquad C = 23{,}0 + \frac{24308 - 19837}{7156} = 23{,}625$ Jahre

5.
$$\begin{array}{rrr} 6703 & 453 & -1073 \\ 7156 & -620 & \\ 6536 & & \end{array}$$

$$D = 23 + \frac{453}{1073} = 23{,}422 \text{ Jahre.}$$

6. $M - D = 0{,}478, \qquad M - C = 0{,}275, \qquad 3(M - C) = 0{,}825$

7. $\Sigma_0 + 3\Sigma_1 + 2\Sigma_2 = 478335, \qquad m^2 = 9{,}839$

$$\mu = \sqrt{9{,}839 - (0{,}400)^2} = 3{,}111 \text{ Jahre}$$

8. $\Sigma_0 + \Sigma_1 = 111949; \qquad \vartheta = \frac{111949 + 2148}{48616} = 2{,}347$ Jahre

9. $\frac{N}{4} = 12154; \qquad Q_1 = 21 + \frac{12154 - 7375}{5759} = 21{,}830$ Jahre

$$Q_3 = 26 - \frac{12154 - 9838}{5249} = 25{,}559 \text{ Jahre}$$

$$Q = 1{,}865 \text{ Jahre}$$

$C - Q_1 = 1{,}795, \qquad Q_3 - C = 1{,}934$

10. $V = 100 \frac{3{,}111}{23{,}900} = 13{,}017$

11. Schiefe $= \frac{0{,}478}{3{,}111} = 0{,}154.$

Tab. 47. Alter der eheschließenden Frauen beim Heiratsalter des Mannes von 25 bis 26 Jahren.[1])

Alter in Jahren X	Zahl der heiratenden Frauen z	s		s'
15—16	4	4		4
16—17	91	95		99
17—18	365	460		559
18—19	1056	1516		2075
19—20	2306	3822		5897
20—21	3553	7375		13272
21—22	5759	13134		21906
22—23	6703	19837	26406	
23—24	7156	26993		
24—25	6536	21623	44083	
25—26	5249	15087		90798
26—27	3500	9838		28996
27—28	2249	6338		19158
28—29	1370	4089		12820
29—30	928	2719		8731
30—31	564	1791		6012
31—32	394	1227		4221
32—33	246	833		2994
33—34	169	587		2161
34—35	109	418		1574
35—36	81	309		1156
36—37	54	228		847
37—38	54	174		619
38—39	28	120		445
39—40	28	92		325
40—41	15	64		233
41—42	11	49		169
42—43	10	38		120
43—44	9	28		82
44—45	6	19		54
45—46	4	13		35
46—47	3	9		22
47—48	2	6		13
48—49	2	4		7
49—50	1	2		3
50—51	1	1		1
	48616			

[1]) Statistik des Deutschen Reichs, Bd. 441, S. 33. Vom Alter von 40 Jahren ab sind die Zahlen für die Altersgruppen durch Interpolation auf graphischem Wege in die Zahlen für die Altersjahre zerlegt worden.

Hervorzuheben sind der große Variationskoeffizient, die positive Schiefe und der Umstand, daß die Hälfte der Bräute dem engen Altersintervall 21,830 bis 25,559 von nur 3,729 Jahren entstammt, während sich die Alter der übrigen über 31 Jahre erstrecken.

Die Altersverhältnisse der Heiratenden hängen im großen hauptsächlich von vier mächtigen Faktoren ab: vom Triebleben, von der Volkstradition, von der wirtschaftlichen Lage und von dem wirtschaftlichen Gefüge. Die Unterschiede, die sich darin bei verschiedenen Völkern zeigen, haben charakteristische Bedeutung, und die Veränderungen, die sich bei einer und derselben Bevölkerung äußern, werden zumeist auf eine Änderung der wirtschaftlichen Lage hinweisen, da die andern Faktoren zeitlichen Änderungen nur in sehr geringem Maße unterworfen sein dürften. Linksseitige Asymmetrie der Verteilung kann a priori vorausgesagt werden, da die stärkste Heiratsfrequenz bei beiden Geschlechtern in die jungen Alter fällt.

Die umstehenden Daten sind geeignet, eine Vorstellung der einschlägigen Verhältnisse zu geben. Sie betreffen die im Jahresdurchschnitt 1910/11 im Deutschen Reich[1]) und in Sachsen[2]) geschlossenen Ehen. Siehe Tab. 48, S. 156.

Nachstehend sind die daraus gezogenen Hauptergebnisse zusammengestellt.

	Deutsches Reich		Sachsen	
	Bräutigam	Braut	Bräutigam	Braut
M	27,412	24,762	26,333	24,361
C	26,379	23,833	25,277	23,507
D	25,033	22,699	23,985	22,664
μ	4,56	4,68	4,14	4,18
Sch.	0,522	0,441	0,567	0,406

Aus der Vergleichung der drei Mittelwerte geht hervor, daß in Sachsen das Heiraten bei beiden Geschlechtern etwas früher erfolgt als im Deutschen Reich im ganzen. Dies hat seinen Grund in der wirtschaftlichen Struktur Sachsens. Sachsen ist bekanntlich das industriereichste Land Deutschlands, ja der Erde überhaupt. Der Industriearbeiter heiratet dann, wenn er eine einigermaßen einträgliche Stellung erlangt hat. Dies ist in wirtschaftlich normalen Zeiten im allgemeinen bereits in jungen Jahren der Fall. In den Agrargebieten des Deutschen Reichs erfolgt dagegen die Eheschließung in der Regel erst dann, wenn der väterliche Hof oder ein anderer landwirtschaftlicher Betrieb zur eigenen Bewirtschaftung übernommen wird. Dies ist fast immer erst in späteren Jahren möglich.

Die Differenz zwischen dem mittleren Heiratsalter des Mannes und dem der Frau ist in Sachsen etwas kleiner als im Reich. Auch dies ist auf den Einfluß der wirtschaftlichen Struktur zurückzuführen. Das wirtschaftliche Gefüge eines Landes ist im allgemeinen auf das Heiratsalter des Mannes von stärkerem Einfluß als auf das der Frau. Infolgedessen liegt in Industrieländern das Heiratsalter des Mannes verhältnismäßig niedrig. Somit nähert sich das mittlere Heiratsalter des Mannes dem der Frau in Industrieländern mehr als in Agrarländern.

[1]) Statistik des Deutschen Reichs, Bd. 246, S. 36, und Bd. 256, S. 36.

[2]) Zeitschrift des Sächsischen Statistischen Landesamtes, 68. Jahrgang, 1922, S. 116.

Tab. 48. Die erstmalig Heiratenden nach dem Alter.

Alter in Jahren X	Deutsches Reich		Sachsen	
	männlich z	weiblich z	männlich z	weiblich z
15—16	—	23	—	—
16—17	—	1 052	—	9
17—18	—	4 369	—	81
18—19	120	12 557	—	314
19—20	707	25 492	3	876
20—21	2 216	38 873	23	1 621
21—22	13 643	52 412	713	2 426
22—23	31 760	55 902	2 146	2 693
23—24	46 505	54 397	2 679	2 558
24—25	56 512	48 168	2 671	2 102
25—26	56 787	39 890	2 258	1 549
26—27	48 770	31 362	1 724	1 150
27—28	40 769	23 965	1 300	788
28—29	33 154	18 161	981	539
29—30	26 097	13 937	716	418
30—31	20 596	10 510	546	311
31—32	16 330	8 042	417	242
32—33	12 674	6 299	327	177
33—34	10 050	4 971	261	146
34—35	7 824	3 941	200	113
35—36	6 129	3 233	164	93
36—37	4 669	2 653	117	79
37—38	3 680	2 072	97	58
38—39	2 725	1 657	71	49
39—40	2 127	1 275	52	39
40—41	1 752	1 100	49	34
41—42	1 486	947	37	32
42—43	1 159	785	29	24
43—44	936	652	25	17
44—45	785	594	19	18
45—46	615	497	14	18
46—47	529	416	15	13
47—48	451	388	8	10
48—49	386	274	12	9
49—50	307	228	10	6
50—51	269	204	7	6
51—52	193	171	5	5
52—53	179	138	5	3
53—54	144	96	5	3
54—55	116	78	2	2
55—56	86	53	4	1
56—57	80	44	2	1
57—58	70	27	2	2
58—59	53	26	2	—
59—60	55	23	2	1
	453 495	471 954	17 720	18 636

Die Streuung der Altersgliederung der Eheschließenden ist in Sachsen geringer als im Reich. Dies hängt vermutlich wiederum mit dem ausgesprochen industriellen Charakter Sachsens zusammen, der bewirkt, daß sich das Heiraten in Sachsen in einem engeren Altersrahmen vollzieht. Sowohl im Deutschen Reich als auch in Sachsen ist die Streuung beim weiblichen Geschlecht größer als beim männlichen. Der Grund hierfür liegt darin, daß die weiblichen Personen früher heiraten als die männlichen.

Die Schiefe der Verteilung ist in Sachsen bei den männlichen Personen größer und bei den weiblichen Personen kleiner als im Reich.

Nach dem Kriege treten die eben angeführten Regelmäßigkeiten nicht mehr so scharf in die Erscheinung. Dies hängt damit zusammen, daß das Land Sachsen infolge seines Industriereichtums unter den ungünstigen wirtschaftlichen Verhältnissen bis 1932 außerordentlich zu leiden hatte. In wirtschaftlich ungünstigen Zeiten liegt im allgemeinen das mittlere Heiratsalter höher als in wirtschaftlich günstigen Zeiten. Nach dem Kriege wirkte somit in Sachsen der konjunkturelle Einfluß in entgegengesetzter Richtung wie der strukturelle. Auf Grund dieser Erwägung ist in Tab. 48 das eheschließungsstatistische Material von 1910/11 herangezogen worden.

Im Anschluß hieran sei noch kurz darauf hingewiesen, daß der Wert des mittleren Heiratsalters von der Altersgliederung der heiratsfähigen Bevölkerung abhängig ist. Sind in der Gesamtheit der heiratsfähigen Bevölkerung die jüngeren Altersklassen stark vertreten, so liegt im allgemeinen das mittlere Heiratsalter niedriger als in einer Bevölkerung, in der diese Altersklassen schwach besetzt sind. Verwendet man zur Ausschaltung dieses Einflusses der Altersgliederung die Heiratstafelmethode, so erhält man für die Jahre 1910/11 die folgenden Werte für das mittlere Alter der erstmalig Heiratenden:

	Männlich	Weiblich
Deutsches Reich[1] .	27,77	25,11
Sachsen[2]..........	26,68	24,99

Die tatsächlichen Mittelwerte liegen unter diesen Tafelmittelwerten. Hieraus folgt, daß sich 1910/11 die heiratsfähige Bevölkerung im progressiven Zustand befand.

§ 5. Korrelation zwischen zwei Variablen.

Theorie.

70. Solange es sich um ein einzelnes Merkmal handelt, richtet sich die Frage nach seiner Verteilung auf die Glieder des Kollektivs; zu ihrer Beschreibung dienen die verschiedenen Mittelwerte, Streuungsmaße und aus ihnen abgeleitete Größen, zu ihrer Veranschaulichung das Häufigkeitspolygon oder die Häufigkeitskurve.

Eine neue Fragestellung ergibt sich, sobald z w e i Merkmale in Betracht kommen, die nicht getrennt, sondern in ihrem vereinigten Auftreten an den Kollektivgliedern verfolgt werden sollen. Man hat es dann mit z w e i V a r i a b l e n

[1] Statistik des Deutschen Reichs, Bd. 275, 1918, S. 40*.
[2] Zeitschrift des Sächsischen Statistischen Landesamtes 1928/29, S. 119.

zu tun, deren durch Zählung oder Messung festgestellte Einzelwerte durch das Kollektiv zu Paaren verbunden sind, so daß jedem Glied ein solches Wertepaar von X, Y zugeordnet ist[1]).

Es gilt gewissermaßen als Axiom, daß zwischen den Merkmalen von Naturgegenständen einer bestimmten Art Menschen, Tieren, Pflanzen, selbst auch von manchen Kunstprodukten, eine gewisse Abhängigkeit besteht, daß sich mit andern Worten eine Ebenmäßigkeit an ihnen ausbildet oder auszubilden strebt, für die wir auf Grund vielfacher Erfahrung eine mehr oder weniger ausgeprägte Empfindung erlangen; eine starke Abweichung von ihr fällt uns sogleich auf und bei größerer Übung im Beobachten und Schauen entgehen uns auch schwächere Grade nicht. Um nur eines der geläufigsten Beispiele zu geben, werden ein auffallend langer oder kurzer Oberkörper, lange oder kurze Arme, langer oder kurzer

[1]) Es ist für das Folgende von grundlegender Bedeutung, daß korrespondierende Werte der beiden in Betracht gezogenen Merkmale, d. h. solche Werte zueinander in Beziehung gesetzt werden, welche derselben statistischen Einheit (also etwa demselben Individuum, wie Länge und Breite eines Blattes, oder demselben Individuenpaar, wie Alter des Mannes und der Frau bei der Eheschließung u. ä.) angehören.

Es gibt aber Fragestellungen, die eine andere Art der Inbeziehungsetzung erfordern. Auf eine solche Art, die bereits eingehendere Beachtung gefunden hat, soll hier hingewiesen werden.

Man ordne die Werte des Merkmals X, die an den Individuen eines Kollektivs vorkommen, steigend oder fallend, so erhält dadurch jeder Merkmalwert in seiner Ordnungsnummer einen Rang oder Grad.

Das gleiche geschehe mit den Werten des andern Merkmals Y.

Auf diese Weise entstehen zwei gleich umfangreiche Wertfolgen.

Wenn dann zwei korrespondierende Merkmalwerte denselben Grad aufweisen, so sollen sie kograduiert heißen, sofern die beiden Folgen gleichartig — beide steigend oder beide fallend — geordnet sind; hingegen kontragraduiert, wenn die beiden Folgen ungleichartig — die eine steigend, die andere fallend — geordnet sind.

Im allgemeinen werden nur einzelne oder auch gar keine Paare korrespondierender Werte kograduiert oder kontragraduiert sein. Den höchsten Grad erlangt die Kograduation, bzw. die Kontragraduation, wenn alle Paare ko-, bzw. kontragraduiert sind. Der Sachverhalt ist dann der, daß im ersten Falle mit dem Wachsen des einen Merkmals stets auch ein Wachsen des andern erfolgt und daß im zweiten Falle mit dem Wachsen des einen Merkmals ein Abnehmen des andern verbunden ist. Diese Grenzfälle sollen als perfekte Kograduation und perfekte Kontragraduation bezeichnet werden.

Eine solche getrennte Reihung der beiden Merkmalwerte wird beispielsweise am Platze sein, wenn es sich um die Frage handelt, welches der Merkmale die größere Variabilität besitzt; man wird dann aus jeder der beiden Reihen ein Streuungsmaß berechnen und aus diesen Streuungsmaßen die Antwort holen. Nur wenn die in Beziehung zu setzenden Größen X, Y gleichartig sind (Längen, Gewichte), kann unter Umständen nach absoluten Streuungsmaßen geurteilt werden; in der Regel aber — bei ungleichartigen Größen immer — werden aber relative Streuungsmaße, wie etwa die Variabilitätskoeffizienten (Art. 67) in Frage kommen.

Auch auf andere Fragen kann das Studium des Verlaufs der beiden Wertreihen Auskunft geben.

Man vergleiche hierzu die Arbeit von C. Gini, Delle relazioni tra le intensità cograduate di due caratteri. Atti del Reale Ist. Veneto, 1916—1917, t. LXXVI, p. 1147—1185.

Hals, großer oder kleiner Kopf von jedem gleich bemerkt; der bildende Künstler, der sich die „normalen Verhältnisse" durch vieles Schauen und Vergleichen eingeprägt hat, nimmt auch eine leise Abweichung von ihnen wahr.

Indessen, die vollkommenste Art der Abhängigkeit, die funktionale, derzufolge jedem Werte der einen Variablen ein bestimmter Wert der andern zugeordnet ist, kommt in den Kollektiven, die den Gegenstand unserer Untersuchungen bilden, nicht vor. Eine andere Art, die im Gegensatze zur funktionalen als korrelative Abhängigkeit bezeichnet werden soll, tritt hier in die Erscheinung. Ihr Wesen besteht darin, daß zwar zu einem Werte von X verschiedene Werte von Y gehören, daß sie aber eine bestimmte durch X bedingte Verteilung aufweisen. Diese Verteilung mit den sie kennzeichnenden Größen, Mittelwerten und Streuungen, ist es, was dem Ausgangswerte X zuzuordnen ist. In ebensolcher Weise ist einem Ausgangswerte von Y eine Verteilung der Werte von X zugeordnet[1]).

So wichtig sich die Auffindung funktionaler Abhängigkeiten für die Erkenntnis und Beherrschung der Natur erwiesen hat, so bedeutungsvoll wird sich die fortschreitende Aufdeckung korrelativer Zusammenhänge für die Erforschung jener zahlreichen Materien erweisen, die der kollektiven Behandlung zu überantworten sind. Welcher große Unterschied zwischen den in weiten Grenzen unbestimmten Angaben über Größen und Größenverhältnisse, mit welchen sich die beschreibenden Naturwissenschaften begnügten und begnügen mußten, und den genauen zahlenmäßigen Darstellungen, welche die statistische Methode liefert! Erst auf dieser Grundlage können Fragen der Entwicklungslehre, der Erblichkeitslehre, der Rassenhygiene, der Krankheitsforschung, der Züchtung von Tieren und Pflanzen erfolgreich in Angriff genommen werden.

An die Stelle der bisher betrachteten reihenförmigen Anordnung der Argumentwerte mit ihren Häufigkeiten kommt jetzt eine flächenhafte Anordnung der Wertverbindungen von X, Y und ihrer Häufigkeiten. Es entsteht so eine Tafel mit zwei Eingängen, geometrisch gesprochen ein Netz, gebildet von zwei Scharen rechtwinklig sich schneidender Parallelen, das quadratisch wird, wenn man nicht die wirklichen Maße, sondern die Klassen zur Grundlage nimmt. Bezüglich der Ausfüllung einer solchen Korrelationstabelle sind zwei verschiedene Fälle zu betrachten.

Die Argumente X, Y nehmen nur bestimmte ganzzahlige Werte an, die durch Zählung festgestellt werden (Organe bei Tieren und Pflanzen, z. B. Drüsen, Staubfäden, Blätter, Samen u. a.): Dann stehen die Häufigkeitszahlen an den Gitterpunkten des Netzes und geben an, wie viele Glieder des Kollektivs die betreffende Wertverbindung an sich tragen.

Die Argumente X, Y sind stetige Variable, ihre Feststellung geschieht durch Messung im weitesten Sinne (Dimensionen von Organpaaren, zwei Dimensionen desselben Organs, Alter zusammengehöriger Personen u. a.): Dann sind die Häufigkeitszahlen in die Felder des Netzes eingetragen und sagen aus, bei wie vielen Gliedern des Kollektivs die Argumente X, Y in die durch das Feld bezeichneten Klassen fallen.

[1]) Für den Begriff des korrelativen Zusammenhangs haben L. v. Bortkiewicz (Die Iterationen. Berlin 1917, S. 3 u. f.) und A. A. Tschuprow (Grundbegriffe und Grundprobleme der Korrelationstheorie. Leipzig 1925, S. 20 u. f.) den Begriff der Stochastik oder der stochastischen Verbundenheit eingeführt. Vgl. hierzu auch A. Timpe, Einführung in die Finanz- und Wirtschaftsmathematik. Berlin 1934, S. 182 und H. Münzner, Grundbegriffe und Probleme der Korrelationsrechnung. Deutsche Mathematik 1936, S. 290 u. f.

Die Zahlen einer Korrelationstabelle ordnen sich in horizontale Reihen oder Zeilen und in vertikale Reihen oder Kolonnen; für beide zusammen soll der Name „Reihe“ gelten. Im allgemeinen steht es frei, welches Merkmal man mit X und welches mit Y bezeichnet. Dem X soll der obere horizontale, dem Y der linke vertikale Rand der Tabelle eingeräumt sein; diese Ränder also bilden das (ursprüngliche) Koordinatensystem. Die Art der Beschreibung einer solchen Tabelle wird aus den später folgenden Beispielen ersichtlich sein.

71. Die Ausfüllung einer Korrelationstabelle geschieht auf Grund der Urliste, nachdem das Netz entworfen und entsprechend beziffert ist. Man nimmt ein Glied nach dem andern und bildet es gemäß der ihm zugehörigen zwei Argumentwerte bei dem entsprechenden Gitterpunkt, bzw. in dem entsprechenden Felde, in irgend einer Weise, durch einen Punkt oder einen Strich o. dgl. ab und zählt nach Erschöpfung der Glieder die so gesetzten Zeichen. Bei sehr umfangreichen Kollektiven kann diese Art beschwerlich und wegen Mangels einer Kontrolle — es bliebe nur eine Wiederholung des ganzen Vorgangs übrig — unsicher werden; man fertigt dann für jedes Glied eine besondere Karte an, auf der die Argumentwerte in einheitlicher Weise aufgezeichnet sind, ordnet die Karten nach den Werten oder Klassen von X in Päckchen, innerhalb jedes Päckchens wieder ordnet man nach Y und erhält schließlich so viele Kartenpäckchen, als es besetzte Wertverbindungen von X, Y gibt, und nimmt nach sorgfältiger Überprüfung die Zählung und Eintragung vor.

Noch ist eine Vereinbarung darüber notwendig, wie man mit der Einreihung solcher Glieder verfährt, bei denen das eine — stetige — Argument mit einer Klassengrenze zusammenfällt, und wie mit solchen, bei welchen beide Argumente auf Klassengrenzen zu liegen kommen. Im ersten Falle ist es üblich, das betreffende Glied je zur Hälfte den beiden an der Grenze zusammenstoßenden Feldern zuzuzählen; im zweiten Falle weist man jedem der beteiligten vier Felder ein Viertel zu. Auf diese Weise können auch gebrochene Häufigkeitszahlen mit den Anhängen 0,25, 0,5, 0,75 zustandekommen. Nicht leicht wird sich die Klasseneinteilung so treffen lassen, daß gebrochene Häufigkeitszahlen vermieden werden.

Die Reihensummen geben die Verteilung des Kollektivs nach je einer Eigenschaft, und zwar liefert die untere Summenreihe die Verteilung des Merkmals X, die rechte Summenreihe die Verteilung des Merkmals Y. Die gemeinsame Summe dieser beiden Summenreihen, in der rechten untern Ecke der Tafel verzeichnet, gibt den Umfang des Kollektivs.

Damit sind die Herstellung und das äußere Bild einer Korrelationstabelle hinreichend beschrieben. Die folgenden Beispiele werden die Vorstellung ergänzen.

72. Beispiele. 1) Zuerst zwei kleine Beispiele aus der biologischen Statistik, betreffend Organe von Trientalis europaea; die Exemplare waren in der Zeit vom 8. bis 23. Juni 1912 am Ufer eines kleinen Waldsees in der Nähe von Lund gesammelt[1]). Alles übrige ist der Beschreibung zu entnehmen. Die erste Tabelle hat es mit zwei unstetigen, die zweite mit einer unstetigen und einer stetigen Variablen zu tun.

---

[1]) C. V. L. Charlier, A Statistical Description of Trientalis europaea, Arkiv för Botanik, Bd. XII, Nr. 14, 1913, S. 6 und 11. Vgl. E. Weber, Einführung in die Variations- und Erblichkeits-Statistik, München 1935, S. 102.

Korrelation einerseits zwischen der Zahl der Blütenstengel und der Zahl der Blumenblätter (Tab. 49), anderseits zwischen der Zahl der Blumenblätter und der Länge des längsten Blumenblattes bei Trientalis europaea (Tab. 50).

Tab. 49.

Zahl der Blumenblätter	Zahl der Blütenstengel			Summe
	1	2	3	
5	119	6	—	125
6	103	51	1	155
7	10	16	2	28
8	1	5	5	11
9	—	—	2	2
	233	78	10	321

Tab. 50.

Länge des längsten Blumenblattes in mm	Zahl der Blumenblätter					Summe
	5	6	7	8	9	
10	1	—	—	—	—	1
15	4	1	—	—	—	5
20	13	4	—	—	—	17
25	24	11	1	—	—	36
30	30	21	2	1	—	54
35	21	26	4	1	—	52
40	17	30	—	3	—	50
45	6	18	6	2	—	32
50	2	15	5	1	1	24
55	1	3	3	—	—	7
60	—	2	2	1	—	5
65	—	2	1	1	1	5
70	—	—	1	1	—	2
	119	133	25	11	2	290

Den Tabellen läßt sich der Hauptsache nach folgendes entnehmen.

Die vorherrschende Zahl der Blütenstengel ist 1, ihre relative Häufigkeit (233 bez. auf die Gesamtzahl 321) beträgt 0,726; die vorherrschende Zahl der Blumenblätter ist 6, ihre relative Häufigkeit 0,459; mit der Zahl der Blütenstengel wächst die Zahl der Blumenblätter, denn die aus den drei Kolonnen abgeleiteten arithmetischen Mittel sind der Reihenfolge nach 5,5, 6,3, 7,8.

Die Hauptmasse der längsten Blütenblätter — 53,8% — fällt in das Intervall von 27,5 bis 42,5 mm. Mit der Zahl der Blumenblätter wächst auch die Länge des längsten unter ihnen; denn die aus den fünf Kolonnen abgeleiteten arithmetischen Mittel 30,8, 38,2, 46,6, 47,3, 57,5 mm zeigen ein ununterbrochenes erhebliches Ansteigen.

2) Ein größeres Beispiel mit unstetigen Variablen liegt der folgenden Korrelationstabelle zugrunde, über deren Inhalt die Überschrift und die ihr folgenden Ausführungen Aufschluß geben.

Tab. 51. Korrelation zwischen der ehelichen Fruchtbarkeit der Väter und ihrer Söhne.

Zahl der Kinder des Sohnes Y	Zahl der Kinder des Vaters X																	Summe
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
0	5	8	7	14	18	2	2	3	8	3	4	4	.	.	.	.	.	78
1	3	3	6	5	8	8	6	5	4	.	2	.	1	.	.	.	.	51
2	7	5	6	13	12	12	12	6	5	4	2	1	1	.	.	.	.	86
3	5	10	13	11	17	13	13	12	10	4	1	2	1	.	.	.	.	112
4	4	16	18	24	23	5	18	10	7	8	1	5	2	1	.	.	1	143
5	9	8	11	14	16	12	16	12	2	5	8	6	3	.	2	.	.	124
6	3	4	10	16	13	11	11	10	10	1	2	2	1	.	.	.	.	94
7	5	6	8	7	10	14	11	12	4	7	1	1	.	.	.	.	.	86
8	3	5	4	15	19	7	10	8	4	2	2	1	.	.	1	.	.	81
9	1	6	5	9	5	5	8	5	4	3	3	1	.	.	.	.	.	55
10	2	3	9	3	2	5	4	6	4	3	1	1	1	.	.	.	.	44
11	1	1	1	4	2	1	1	2	1	1	.		.	.	.	.	.	15
12	.	.	2	2	2	.	1	1	1	2	.	1	.	.	.	.	.	12
13	1	.	.	1	3	.	1	1	1	1	.	1	.	1	.	.	.	11
14	.	1	.	.	.	1	.	.	.	.	1	.	.	.	.	.	.	3
15	.	.	1	.	.	.	.	1	.	.	1	.	.	.	.	.	.	3
16		.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	2
Summe	49	76	101	138	150	96	114	94	66	45	29	26	10	2	3	.	1	1000

Die Ehe des Sohnes war entweder beendet durch seinen oder der Gattin Tod oder sie hatte mindestens 15 Jahre gedauert zur Zeit der Erhebung. Auf die Dauer der Elternehe wurde keine Rücksicht genommen. Aus jeder Familie wurde nur ein Sohn genommen. Das Material stammt aus englischen Baronets- und Peersfamilien[1].

Als Glied des Kollektivs tritt hier ein Paar auf, bestehend aus Vater und Sohn. Die häufigste Kinderzahl des Vaters ist 5, die des Sohnes 4. Um den Punkt 5/4 drängen sich auch die größten Häufigkeitszahlen. Im übrigen ist der Verlauf dieser Zahlen ein recht unregelmäßiger, was mit dem kleinen Umfang des Kollektivs zusammenhängt; die 1000 Glieder verteilen sich auf 176 besetzte Wertverbindungen. Das am stärksten besetzte Feld ist 4/4.

Bei der Beurteilung der Tatsache, daß die häufigsten Kinderzahlen bei Vätern und Söhnen nicht übereinstimmen, muß darauf geachtet werden, daß bei den Söhnen auch kinderlos gebliebene gezählt werden, was bei den Vätern ausgeschlossen ist.

3) Es folgen zwei kleine Beispiele rein stetiger Kollektive aus der biologischen Statistik, wieder auf *Trientalis europaea* bezüglich wie das erste Beispiel[2]). Die Überschriften reichen zum Verständnis der Tabellen hin.

---

[1]) K. Pearson, A. Lee, L. Bramley-Moore, Mathematical Contributions to the Theory of Evolution. Phil. Trans. Roy. Soc., A, vol. 192 (1899), p. 287, 321.

[2]) C. V. L. Charlier, A Statistical Description of Trientalis europaea. Arkiv för Botanik, Bd. XII, Nr. 14, 1913, S. 22, 24.

Tab. 52. Korrelation zwischen der Stammdicke und der Länge des längsten Blumenblattes bei Trientalis europaea.

Länge des längsten Blumenblattes (in mm)	Stammdicke (in mm)											Summe
	0,425	0,525	0,625	0,725	0,825	0,925	1,025	1,125	1,225	1,325	1,425	
10,5	1	.	.	.	.	.	.	.	.	.	.	1
16,5	1	4	1	1	.	.	.	.	.	.	.	7
22,5	1	9	16	3	1	.	.	.	.	.	.	30
28,5	.	2	9	22	9	2	1	.	.	.	.	45
34,5	.	.	8	19	20	4	1	.	.	.	.	52
40,5	1	.	.	7	18	12	6	4	.	.	.	48
46,5	.	.	.	1	8	9	3	2	1	.	.	24
52,5	.	.	.	.	.	3	6	4	1	.	.	14
58,5	.	.	.	.	.	.	2	2	1	2	.	7
64,5	.	.	.	.	.	.	.	.	1	3	.	4
70,5	.	.	.	.	.	.	.	.	1	.	1	2
Summe	4	15	34	53	56	30	19	12	5	5	1	234

Tab. 53. Korrelation zwischen der Breite und der Länge des längsten Blumenblattes bei Trientalis europaea.

Länge (in mm)	Breite (in mm)												Summe
	5,5	7,5	9,5	11,5	13,5	15,5	17,5	19,5	21,5	23,5	25,5	27,5	
10	1	.	.	.	.	.	.	.	.	.	.	.	1
15	1	3	.	.	.	.	.	.	.	.	.	.	4
20	1	2	7	.	.	.	.	.	.	.	.	.	10
25	.	4	14	6	.	.	.	.	.	.	.	.	24
30	.	.	2	17	3	3	1	.	.	.	.	.	26
35	.	.	.	2	11	9	1	.	.	.	.	.	23
40	.	.	.	.	10	8	7	1	.	.	.	.	26
45	.	.	.	.	1	8	5	2	1	1	1	.	19
50	.	.	.	.	.	2	1	5	1	2	.	.	11
55	.	.	.	.	.	.	.	.	1	.	.	.	1
60	.	.	.	.	.	.	.	3	.	.	1	.	4
65	.	.	.	.	.	.	.	.	.	.	.	1	1
70	.	.	.	.	.	.	.	.	.	1	1	.	2
Summe	3	9	23	25	25	30	15	11	3	4	3	1	152

Es ist durchwegs die Klassenmitte angegeben. Die Tabellen bieten ein von den früheren verschiedenes Bild dar, das sich dadurch auszeichnet, daß sich die besetzten Klassen deutlich um eine Gerade, u. zw um eine Diagonale der Tabelle zusammenscharen.

4) Das folgende größere Beispiel eines stetigen Kollektivs gehört wieder in das Gebiet der Pflanzenbiologie und zeigt, wie sich beim wilden Efeu (Hedera helix) Länge und Breite der Blätter verteilen. Die Einheit, in der beide ausgedrückt sind, beträgt 1/8 englische Zoll und die Klassengröße, ebenfalls gemeinsam, ist zwei solche Einheiten[1]).

Tab. 54. Korrelation zwischen der Länge und Breite der Blätter von Hedera helix.

Breite Y	Länge X													Summe
	2,95	4,95	6,95	8,95	10,95	12,95	14,95	16,95	18,95	20,95	22,95	24,95	26,95	
2,95	7	13	.	.	.	.	.	.	.	.	.	.		20
4,95	26	97	18	1	.	.	.	.	.	.	.	.		142
6,95	3	101	106	37	4	.	.	.	.	.	.	.		251
8,95	3	33	190	152	31	4	2	.	.	.	.	.		415
10,95	.	7	88	227	98	16	2	.	1	.	.	.		439
12,95	.	1	26	137	216	66	7	.	.	.	.	.		453
14,95	.	.	.	55	136	104	22	6	1	.	.	.		324
16,95	.	.	3	11	50	89	49	9	2	.	.	.		213
18,95	.	.	.	4	17	43	31	11	5	1	.	.		112
20,95	.	.	.	.	4	9	21	17	4	1	1	.		57
22,95	.	.	.	.	1	.	3	9	6	1	.	1		21
24,95	.	.	.	.	.	.	4	11	7	5	1	.		28
26,95	.	.	.	.	.	.	.	4	2	4	2	.		12
28,95	.	.	.	.	.	.		1	2	1	3	.		7
30,95	.	.	.	.	.	.	.	.	1	2	.	.		3
32,95	.	.	.	.	.	.	.	.	.	.	2	.		2
34,95	.	.	.	.	.	.	.	.	.	.	1	.		1
36,95														
Summe	39	252	431	624	557	331	141	68	31	15	10	1		2500

[1]) K. Pearson, Mathematical Contributions to the Theory of Evolution. Phil. Trans. Roy. Soc., A, vol. 197 (1901), p. 346, 353.

Die Blätter waren 100 Pflanzen von zwei verschiedenen Lokalitäten, je 25 von jeder Pflanze, entnommen. Als Breite wurde der Abstand jener Tangenten gemessen, die der Längenausdehnung parallel laufen.

Die am häufigsten vorkommende Länge ist 9,95, die Breite 13,95. Die größte Feldhäufigkeit kommt der Wertverbindung 9,95/11,95 zu, um sie scharen sich die größten Häufigkeiten.

73. Um eine zweifach ausgedehnte Verteilung geometrisch darzustellen, muß man aus der Ebene in den Raum hinaustreten, indem man in den Gitterpunkten, bzw. in den Feldmitten der Korrelationstafel auf ihrer Ebene Senkrechte errichtet und auf diesen die Häufigkeiten nach einem passend gewählten Maßstab aufträgt. Werden die Strecken durch Stifte versinnlicht, so geben deren obere Enden eine Vorstellung von der Verteilung. Bei einem unstetigen Kollektiv hätte es eigentlich bei diesem Punktebild zu verbleiben. Wenn hingegen bei einem stetigen Kollektiv der Umfang beständig wächst, die Größe der Felder abnimmt, ohne daß die Häufigkeiten aufhören, endliche, von Null verschiedene Zahlen zu sein, dann verdichtet sich das Punktebild und läßt immer deutlicher eine krumme Fläche in die Erscheinung treten, die man, als das der Häufigkeitskurve entsprechende räumliche Gebilde, *Häufigkeitsfläche* nennen wird. Nimmt man das ganze unter ihr liegende Volumen als Volumeinheit, so gibt der über einem Teil der Basis errichtete Zylinder, mit dieser Einheit gemessen, die *relative* Häufigkeit solcher Glieder des Kollektivs, deren Argumente sich innerhalb der durch die Grundfläche bestimmten Grenzen halten.

Eine andere körperliche Darstellung besteht darin, daß man über den Feldern der Tabelle Prismen von der Höhe der Häufigkeit, diese nach einem passenden Maßstabe aufgetragen, errichtet. Ein solches, dem Staffelpolygon nachgebildetes *Stereogramm* gibt gleichfalls eine anschauliche Vorstellung. Eine zeichnerische Darstellung davon ist aber nur ein unvollkommenes Mittel der Versinnlichung, weil sie kompliziert ausfällt und immer nur einen Teil überblicken läßt.

Was von den Verteilungspolygonen, den Staffelbildern und den Häufigkeitskurven gesagt worden, gilt in verstärktem Maße von den Stereogrammen und Häufigkeitsflächen; sie bieten sehr mannigfache Formen dar, die nur dann als kennzeichnend für eine Materie gelten können, wenn sie sich bei an verschiedenen Materien derselben Art wiederholten Untersuchungen immer wieder einstellen.

Als Vergleichsfläche kann die *normale Häufigkeitsfläche* dienen, jene Fläche nämlich, die durch Drehung der normalen Häufigkeitskurve[1]) um ihre Symmetrieachse entsteht. Bezogen auf ein Koordinatensystem, welches die Drehachse zur z-Achse hat, während die beiden andern Achsen den bisher benützten Achsenrichtungen parallel sind, hat diese Fläche die Gleichung

$$z = \frac{h^2}{\pi} e^{-h^2(x^2+y^2)}. \qquad (1)$$

Nicht nur alle Meridiane, sondern auch alle zur z-Achse parallelen Schnitte derselben sind normale Häufigkeitskurven mit demselben Parameter; es genügt, um dies einzusehen, einen Schnitt parallel zur zx-Ebene, also mit konstantem y zu betrachten; er hat in seiner Ebene eine Gleichung von der Form

$$z = C e^{-h^2 x^2}, \qquad (2)$$

wenn C für $\frac{h^2}{\pi} e^{-h^2 y^2}$ geschrieben wird.

[1]) Vgl. hierzu Art. 117.

So wie eine Verteilung, die sich der normalen Häufigkeitskurve genau anpaßt, sehr selten oder streng genommen niemals anzutreffen sein wird, so wird sich auch kein mit zwei Variablen ausgestattetes Kollektiv ausfindig machen lassen, das zu einer normalen Häufigkeitsfläche führt; nur angenähert trifft dies bei manchen Materien zu. Im übrigen wird diese Angelegenheit an einer späteren Stelle eingehender zur Sprache gebracht werden (III. Abschnitt, § 4).

74. Wir kehren jetzt zur weiteren rechnerischen Bearbeitung einer Korrelationstabelle zurück.

Der erste Schritt besteht darin, daß man auf die einzelnen Reihen, also Zeilen und Kolonnen, und auf die beiden Summenreihen jene Methoden anwendet, die für Verteilungen einer Variablen ausgebildet worden sind. Man wird sich, wenigstens zunächst, darauf beschränken, jede solche Reihe durch ihr arithmetisches Mittel und ihre mittlere Abweichung zu charakterisieren.

Es ergeben sich demnach

1) so viele Zeilenmittel $M_X^{(Y)}$, als es Zeilen gibt, jedes zu einem bestimmten Y gehörig, und die entsprechenden mittleren Abweichungen $\mu_X^{(Y)}$;

2) so viele Kolonnenmittel $M_Y^{(X)}$, als es Kolonnen gibt, mit den zugehörigen mittleren Abweichungen $\mu_Y^{(X)}$;

3) der Mittelwert aller X, M_X, und die entsprechende mittlere Abweichung μ_X;

4) der Mittelwert aller Y, M_Y, und die entsprechende mittlere Abweichung μ_Y.

Bis zu diesem Punkte ist das folgende

Beispiel geführt, das ein Seitenstück bildet zu der Tabelle 51, Art. 72; hier handelt es sich um die Vererbung der mütterlichen Fruchtbarkeit auf die Tochter. Das Material[1]) stammt wie dort aus englischen Baronets- und Peersfamilien und bei seiner Zusammentragung wurde an dem Grundsatz festgehalten, daß die Ehe beiderseits mindestens 15 Jahre gedauert haben muß; aus jeder Ehe wurde nur eine Tochter zufällig herausgegriffen, wo deren mehrere waren. Das Bild der Tabelle ist insofern verändert, als die Felder, wie es nach der Gleichwertigkeit der Klassen sein sollte, nicht quadratisch sind — aus Raumrücksichten.

Die Reihe der Werte $M_Y^{(X)}$ zeigt, daß die Fruchtbarkeit der Töchter mit jener der Mütter wächst; dieser Sachverhalt hält an bis zu den Müttern mit 11 Kindern; von da ab sind die Resultate unregelmäßig und wegen der kleinen Häufigkeiten auch unverläßlich.

Die Reihe der $M_X^{(Y)}$ weist darauf hin, daß auch hohe Fruchtbarkeit der Tochter auf große Fruchtbarkeit der Mutter schließen läßt; bei Töchtern mit 9 und mehr Kindern sind die Ergebnisse minder regelmäßig und verläßlich.

Man kann als Folgerung ableiten, daß die weibliche Fruchtbarkeit ihrem Maße nach eine erbliche Eigenschaft ist.

Der durchwegs erhebliche Unterschied zwischen den $M_Y^{(X)}$ und $M_X^{(Y)}$ erklärt sich daraus, daß unter den Töchtern auch kinderlose vorkommen, was bei den Müttern der Natur der Sache nach ausgeschlossen ist.

[1]) K. Pearson, A. Lee, L. Bramley-Moore, Mathematical Contributions to the Theory of Evolution. Phil. Trans. Roy. Soc., A, vol. 192 (1899), p. 285, 319.

Tab. 55. Korrelation zwischen der ehelichen Fruchtbarkeit der Mütter und ihrer Töchter.

Zahl der Kinder der Tochter Y	Zahl der Kinder der Mutter X																Summe	$M_X^{(Y)}$	$\mu_X^{(Y)}$
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16			
0	5	9	11	18	21	15	8	9	6	3	2	3	.	.	.	.	110	5,39	2,57
1	12	5	14	15	10	13	9	8	5	3	2	2	.	.	.	.	98	5,10	2,78
2	9	9	10	15	18	15	9	3	2	4	2	.	.	.	.	1	97	4,95	2,55
3	5	10	16	11	9	14	13	10	4	8	2	3	.	.	.	.	105	5,63	2,83
4	5	5	19	17	21	15	18	10	14	2	1	5	1	.	.	.	133	5,80	2,64
5	7	6	7	17	23	9	12	13	14	8	3	2	2	.	.	.	123	6,13	2,83
6	4	5	8	11	15	12	15	14	7	5	3	3	1	.	.	.	103	6,22	2,70
7	5	4	3	8	4	13	9	8	5	10	2	1	1	.	.	.	73	6,45	2,91
8	1	2	4	12	9	9	8	5	12	3	4	1	2	1	.	.	73	6,77	2,84
9	.	.	4	3	3	4	7	5	3	2	2	1	.	.	.	.	34	6,85	2,42
10	.	.	1	2	1	3	4	6	3	2	.	1	.	1	.	.	24	7,63	2,43
11	.	.	2	1	1	1	.	.	1	2	.	.	.	.	.	.	8	6,25	2,82
12	.	2	1	2	3	.	1	1	.	.	1	.	1	.	1	.	13	6,46	4,01
13	.	.	.	.	2	1	.	.	.	.	1	.	2	.	.	.	6	8,83	3,58
Summe	53	57	100	132	140	124	113	92	76	52	25	22	10	2	1	1	1000		
$M_Y^{(X)}$	3,15	3,53	3,74	4,02	4,09	4,15	4,69	4,85	5,11	5,13	5,52	4,23	8,10	9,00	12,00	2,00			
$\mu_Y^{(X)}$	2,28	2,79	2,79	2,94	3,03	2,96	2,77	2,95	2,74	2,84	3,43	2,76	3,24	1,00	0	0			

Noch bleiben die allgemeinen Mittel und die mittleren Abweichungen in den Summenreihen anzugeben: sie haben folgende Werte:

$$M_Y = 4{,}34 \qquad M_X = 5{,}90$$
$$\mu_Y = 2{,}97 \qquad \mu_X = 2{,}83$$

Hervorzuheben ist noch, daß sich die mittleren Abweichungen aller Reihen, auch der Summenreihen, auf ziemlich gleicher Höhe halten. Eine Ausnahmestellung nehmen die kinderlos gebliebenen Töchter ein; die analoge Erscheinung zeigt sich in der Männertafel.

Die auf diese letztere bezüglichen Zahlen sind des Vergleiches wegen hierhergesetzt.

X	$M_Y^{(X)}$	$\mu_Y^{(X)}$	Y	$M_X^{(Y)}$	$\mu_X^{(Y)}$
1	4,55	3,09	0	5,54	3,10
2	4,76	3,05	1	5,65	2,65
3	5,12	3,16	2	5,62	2,72
4	4,98	3,09	3	5,68	2,69
5	4,70	3,14	4	5,65	3,08
6	5,11	2,88	5	6,22	3,36
7	5,19	2,72	6	5,87	2,60
8	5,62	3,05	7	5,87	2,64
9	5,00	3,59	8	5,69	2,58
10	5,98	3,59	9	5,93	2,79
11	5,31	3,85	10	6,02	3,08
12	4,96	3,35	11	5,33	2,53
13	4,50	2,33	12	6,67	2,97
14	8,50	4,50	13	7,27	3,62
15	6,00	1,41	14	6,33	3,69
16	—	—	15	7,33	3,31
17	4.00	0.00	16	9,50	0,50

$$M_Y = 5{,}07 \qquad M_X = 5{,}83$$
$$\mu_Y = 3{,}15 \qquad \mu_X = 2{,}90$$

Die Resultate sind minder regelmäßig und wegen der größeren mittleren Abweichungen auch weniger verläßlich; wohl zeigt sich, wenn man die $M_Y^{(X)}$ gruppenweise zu je 3 vereinigt, in den Summen

14,43
14,79
15,81
16,25
19,00

ein Anwachsen, nicht so bei den dreigliedrigen Gruppen aus den $M_X^{(Y)}$:

16,81
17,55
17,43
17,28

Die Erblichkeit der männlichen Fruchtbarkeit ist aus dem vorliegenden Material nicht sicher zu erschließen.

75. Zum Zwecke des weiteren Eindringens in die Sache benützen wir die geometrische Darstellung und beginnen damit, daß wir in dem Koordinatensystem, das wir der Korrelationstabelle zugrunde gelegt haben, die Reihenmittel durch Punkte abbilden. Wir bekommen auf diese Weise zwei Punktreihen, die der Kolonnen- und die der Zeilenmittel.

Denken wir zuerst an den idealen Fall einer normalen Verteilung, so lägen die Kolonnenmittel auf einer Geraden parallel zur X-Achse, die Zeilenmittel auf einer Geraden parallel zur Y-Achse. Ihr Schnittpunkt wäre passend als Mittelpunkt der Korrelationstabelle oder der Korrelation selbst zu bezeichnen, er entspräche dem Gipfelpunkt der Häufigkeitsfläche, der dichtesten Wertverbindung XY. Die beiden Variablen wären in einem solchen Falle voneinander unabhängig in dem Sinne, daß die Mittelwerte der Y für alle X und ebenso die Mittelwerte der X für alle Y gleich wären.

In den praktisch vorkommenden Fällen wird der Sachverhalt fast ausnahmslos ein davon wesentlich verschiedener sein: die Kolonnenmittel werden auf eine im allgemeinen gekrümmte Linie und ebenso die Zeilenmittel auf eine andere krumme Linie hinweisen. Man hätte darauf auszugehen, diese Linie unter möglichster Anpassung an die Punkte analytisch darzustellen, womit man in die Lage versetzt wäre, das zu einem beliebigen X gehörige Kolonnenmittel und ebenso das zu einem beliebigen Y gehörige Zeilenmittel durch bloße Auswertung einer Formel zu finden.

In vielen Fällen jedoch liegen die Punkte so, daß man ihnen eine Gerade anpassen kann, und auf diesen Standpunkt wollen wir uns von jetzt ab stellen.

Wir nehmen also an, die Kolonnenmittel lägen (vorerst in aller Strenge) auf einer Geraden KK', die Zeilenmittel auf einer Geraden ZZ'; den Schnittpunkt M der beiden Geraden nennen wir den Mittelpunkt[1]) der ganzen Verteilung und wollen zuerst nachweisen, daß die durch ihn parallel zu den Koordinatenachsen geführten Geraden auf diesem Punkte M', M'' ausschneiden, welche den Mittelwerten aller Y, beziehungsweise aller X entsprechen, so daß M' die Abbildung von M_Y, M'' die Abbildung von M_X ist, Fig. 21.

Fig. 21. Grundlinien einer linearen Korrelation.

Um dies zu erweisen, bezeichne man die Abweichungen der X von $M''\,M\,P$ mit x, die Abweichungen der Y von $M'\,M\,Q$ mit y, den Richtungskoeffizienten der ZZ' gegen die Vertikale mit b_1, den Richtungskoeffizienten der KK' gegen die Horizontale mit b_2.

Dann ist, weil $M_X^{(Y)}$ den Mittelwert aller X (unter Multiplikation mit der zugehörigen Häufigkeitszahl) längs der Geraden (y) bedeutet,

$$\Sigma^{(y)}(x) = m\, b_1\, y,$$

[1]) Der Punkt M ist der Schwerpunkt der Verteilungstafel. Vgl. P. Riebesell, Die Bedeutung des Korrelationskoeffizienten für Theorie und Praxis der Versicherung. Blätter für Versicherungs-Mathematik, Heft Nr. 3, 1929, S. 117 u. f. und A. Timpe, Einführung in die Finanz- und Wirtschaftsmathematik. Berlin 1934, S. 182 u. f.

wobei sich die Summe auf alle x (unter Multiplikation mit der zugehörigen Häufigkeitszahl) längs der Geraden (y) bezieht; m bedeutet die Summe der Häufigkeitszahlen längs (y). Ebenso besteht

$$\Sigma^{(x)}(y) = n\, b_2\, x,$$

wobei n gleich der Summe der Häufigkeitszahlen längs (x) ist.

Summiert man den ersten Ansatz über alle y, den zweiten über alle x, so kommt man zu den Gleichungen

$$\Sigma\,(n\,x) = b_1\,\Sigma\,(m\,y), \qquad \Sigma\,(m\,y) = b_2\,\Sigma\,(n\,x),$$

wenn man in der ersten Gleichung links zuerst bei konstantem x über alle y summiert, wodurch die Häufigkeit n auftritt, und nachträglich nach x die Summe bildet; analog ist in der zweiten Gleichung zu verfahren.

Die beiden Summen $\Sigma\,(n\,x)$, $\Sigma\,(m\,y)$ genügen also den homogenen linearen Gleichungen

$$\begin{aligned} \Sigma\,(n\,x) - b_1\,\Sigma\,(m\,y) &= 0 \\ b_2\,\Sigma\,(n\,x) - \quad \Sigma\,(m\,y) &= 0, \end{aligned}$$

deren Determinante $1 - b_1\,b_2$ nicht Null ist, wenn, wie wir voraussetzen wollen, die beiden Geraden KK', ZZ' voneinander verschieden sind; infolgedessen führen die Gleichungen zu

$$\Sigma\,(n\,x) = 0, \qquad \Sigma\,(m\,y) = 0$$

und damit ist erwiesen, daß die Punkte M', M'' die Gesamtmittel M_Y, M_X darstellen.

Es handelt sich jetzt darum, die aus der Verteilung resultierenden Werte von b_1 und b_2, also die Geraden KK', ZZ' selbst zu bestimmen, deren Schnittpunkt M bekannt ist, sobald die Gesamtmittel M_X, M_Y gebildet sind. Dazu eignet sich der Mittelwert der Produkte aller Paare zugeordneter Abweichungen x/y, er heiße p; dieser Definition gemäß ist

$$\Sigma\,(x\,y) = N\,p, \tag{3}$$

wenn N den Umfang des Kollektivs bedeutet. Die linksstehende Summe kann auf zwei Arten ausgeführt werden, indem man einmal bei festem y nach x und hierauf nach y summiert oder den umgekehrten Weg einschlägt; auf diese Weise erhält man

$$\begin{aligned} \Sigma\Sigma^{(y)}(x\,y) &= \Sigma\left(y\,\Sigma^{(y)}(x)\right) = b_1\,\Sigma\,(m\,y^2) = N\,b_1\,\mu_Y^2 \\ \Sigma\Sigma^{(x)}(x\,y) &= \Sigma\left(x\,\Sigma^{(x)}(y)\right) = b_2\,\Sigma\,(n\,x^2) = N\,b_2\,\mu_X^2, \end{aligned}$$

wobei μ_Y, μ_X die mittleren Abweichungen in den Summenreihen der Korrelationstabelle bedeuten. Damit aber ergibt sich im Hinblick auf (3)

$$b_1 = \frac{p}{\mu_Y^2} \qquad b_2 = \frac{p}{\mu_X^2}. \tag{4}$$

Unter dem Namen Korrelationskoeffizient ist nun eine neue Größe r eingeführt worden[1]) mittels der Definition

$$r = \frac{p}{\mu_X \mu_Y}, \tag{5}$$

deren Vorzeichen mit dem Vorzeichen von p übereinstimmt, wenn man die Festsetzung trifft, daß die im Nenner auftretenden mittleren Abweichungen mit ihren absoluten Werten genommen werden. Durch diese neue Größe drücken sich b_1, b_2 wie folgt aus:

$$b_1 = r\frac{\mu_X}{\mu_Y} \qquad b_2 = r\frac{\mu_Y}{\mu_X}, \tag{6}$$

und es schreiben sich die Gleichungen der Geraden ZZ', KK', auf den Mittelpunkt M der Korrelation als Ursprung bezogen,

$$\xi = r\frac{\mu_X}{\mu_Y}\eta \qquad \eta = r\frac{\mu_Y}{\mu_X}\xi \tag{7}$$

Die ganze hier vorliegende Angelegenheit läßt noch eine andere Auffassung zu, und ihre Verfolgung wird uns die Bedeutung vorstehender Gleichungen von einer andern Seite näherbringen. Wir hatten bisher vorausgesetzt, daß die Reihenmittel in aller Strenge auf Geraden liegen; das wird nie genau zutreffen, im besten Falle wird es mit einer gewissen Annäherung der Fall sein. Es stellt sich somit die Aufgabe ein, den Zeilenmitteln einerseits und den Kolonnenmitteln anderseits je eine Gerade möglichst anzupassen. Dies soll so geschehen, daß die Quadratsumme der Abweichungen aller mit Häufigkeitszahlen besetzten Punkte der Tabelle, diese Häufigkeitszahlen als Gewichte behandelt, ein Minimum werde; dabei sollen die Abweichungen von ZZ' in Richtung der Zeilen, die Abweichungen von KK' in Richtung der Kolonnen gemessen werden. Schreibt man den supponierten Geraden, von welchen verlangt werden muß, daß sie durch den Punkt M gehen, in Bezug auf ihn als Ursprung die Gleichungen $\xi = b_1\eta$, $\eta = b_2\xi$ zu, so hat ein Punkt x/y von der ersten Geraden die Abweichung $x - b_1 y$, von der zweiten Geraden die Abweichung $y - b_2 x$, und gehört zu ihm die Häufigkeitszahl z, so lauten die Bedingungen:

$$\Sigma[z(x - b_1 y)^2] \text{ ein Minimum}$$
$$\Sigma[z(y - b_2 x)^2] \text{ ein Minimum.}$$

Aus der ersten folgt durch Differentiation nach b_1

$$\Sigma[z(x - b_1 y)y] = 0$$

und weiter bei entsprechender Auflösung unter Benützung der obigen Bezeichnungen

$$\Sigma(xy) - b_1\Sigma(my^2) = 0$$

oder

$$Np - b_1 N\mu_Y^2 = 0,$$

woraus $b_1 = \frac{p}{\mu_Y^2}$; bei Anwendung des gleichen Verfahrens auf die zweite Bedingung ergibt sich aus dieser $b_2 = \frac{p}{\mu_X^2}$.

[1]) Die Einführung der Produktsumme $\Sigma(xy)$ geht auf A. Bravais (Analyse mathématique sur les probabilités des erreurs de situation d'un point, Mém. présentés par divers savants, IIe ser., t. IX [1846], p. 255) zurück. Der Idee nach findet sich der Korrelationskoeffizient zuerst bei F. Galton (insbesondere „Correlations and their Measurment", Proc. Roy. Soc., XIV [1888], p. 135); weiter ausgebildet wurde dieser weittragende Begriff von F. Y. Edgeworth, K. Pearson und G. U. Yule.

Es bestimmen also die Formeln (4) und die aus ihnen abgeleiteten (6) unter allen Umständen jene durch M geführten Geraden, welche sich der Gesamtheit der besetzten Punkte und damit auch den sie vertretenden Zeilen-, bzw. Kolonnenmitteln am besten anpassen, diese Wendung in dem oben dargelegten Sinne verstanden. Wie weit man von diesen Geraden Gebrauch machen kann, d. h. bei welcher Anordnung der Zeilen- und Kolonnenmittel man noch von einer zulässigen Approximierung durch gerade Linien sprechen kann, darüber soll keine allgemeine Entscheidung getroffen werden, das soll vielmehr in jedem einzelnen Falle dem Ermessen überlassen bleiben. Man muß dieses Sachverhalts eingedenk bleiben, wenn man von der vorgeführten Theorie allgemein Gebrauch macht, wie dies tatsächlich in der Regel geschieht. Wir werden später auf Fälle zu sprechen kommen, die ein anderes Verhalten zeigen.

Um die Minima, die das Maß der Annäherung bestimmen, zu erhalten, hat man bei Verfolgung der ersten Summe in deren Entwicklung

$$\Sigma(x^2) - 2b_1\,\Sigma(xy) + b_1^2\,\Sigma(y^2),$$

worin sich die Summen auf alle Werte, bzw. Wertverbindungen von x, y, entsprechend ihren Häufigkeiten, beziehen und daher der Reihe nach durch $N\mu_X^2$, Np, $N\mu_Y^2$ zu ersetzen sind, den Wert von b_1 aus (4) einzutragen und findet so:

$$N\left[\mu_X^2 - 2\frac{p^2}{\mu_Y^2} + \frac{p^2}{\mu_Y^2}\right] = N\left[\mu_X^2 - \frac{p^2}{\mu_Y^2}\right]$$

also mit Rücksicht auf (5)

$$\min \Sigma\,[z(x - b_1 y)^2] = N\mu_X^2(1 - r^2); \qquad (8)$$

die gleiche Behandlung der zweiten Summe führt zu

$$\min \Sigma\,[z(y - b_2 x)^2] = N\mu_Y^2(1 - r^2). \qquad (9)$$

Da es sich um Quadratsummen handelt, müssen die letztgefundenen Ausdrücke positiv sein, woraus hervorgeht, daß 1 die obere Grenze des absoluten Wertes von r ist. Ist sie erreicht, dann hat man, gemäß den Formeln (6), $b_1 b_2 - 1 = 0$, die beiden Geraden ZZ', KK' fallen in eine zusammen. Die Bedeutung hiervon wird noch zu besprechen sein.

Eine gute Kennzeichnung der Art und des Grades der Korrelation gibt der Winkel Θ, den die beiden Mittelachsen ZZ', KK' miteinander bilden; über seine Zählungsweise gibt Fig. 22 Aufschluß. Man hat für ihn die Bestimmung

$$\operatorname{tg}\Theta = \frac{\dfrac{1}{b_2} - b_1}{1 + \dfrac{b_1}{b_2}} = \frac{1 - b_1 b_2}{b_1 + b_2} = \frac{1 - r^2}{\left(\dfrac{\mu_X}{\mu_Y} + \dfrac{\mu_Y}{\mu_X}\right)r}. \qquad (10)$$

Zunächst beachte man, daß das Vorzeichen von $\operatorname{tg}\Theta$ mit dem Vorzeichen von r übereinstimmt, weil $1 - r^2$ und $\dfrac{\mu_X}{\mu_Y} + \dfrac{\mu_Y}{\mu_X}$ positiv sind. Eine positive Korrelation (d. h. eine mit positivem Korrelationskoeffizienten) ist also durch einen spitzen Winkel gekennzeichnet. Das Wesen einer solchen besteht darin, daß mit dem Wachsen der einen Variablen die Mittelwerte der andern wachsen, wie dies durch Fig. 22, α) angedeutet ist.

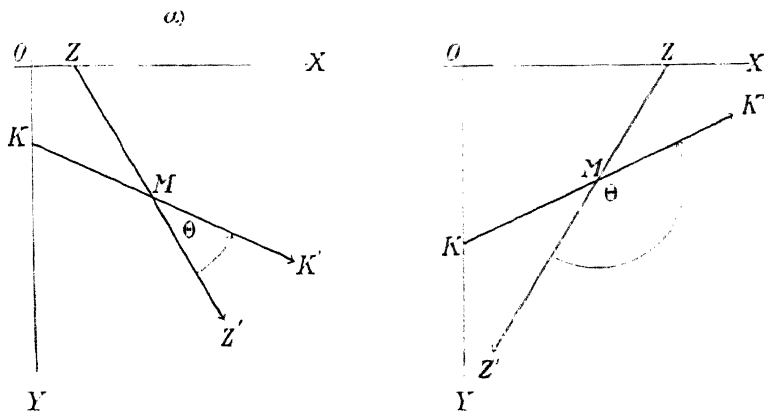



Fig. 22. Positive und negative Korrelation.

Der Fall einer **negativen Korrelation** führt zu einem stumpfen Θ, und ihr Wesen ist dadurch gekennzeichnet, daß dem Wachsen der einen Variablen ein Abnehmen der Mittelwerte der zweiten Variablen zugeordnet ist; man vergleiche dazu Fig. 22, β).

Sind die beiden Variablen unabhängig voneinander, so sind p und r gleichzeitig Null, die Mittelachsen parallel den Koordinatenachsen, Θ ein rechter Winkel. Es wäre aber zu weitgehend, wenn man, weil in einem besondern Falle die Rechnung $r = 0$ ergeben hat, auf Unabhängigkeit schließen wollte; es kann sein, daß trotzdem eine schwache Korrelation mit zu den Achsen sehr wenig geneigten Mittelachsen besteht. Darum ist es vorsichtiger, das Rechnungsergebnis $r = 0$ dahin zu deuten, daß sich die Variablen bei dem vorhandenen Material als **korrelationslos** erweisen.

Der andere Grenzfall, zerfallend in $r = +1$ und $r = -1$, der gleichfalls praktisch kaum jemals vorkommen wird, würde dazu führen, daß die Mittelachsen in **eine** Gerade zusammenfallen; man bezeichnet ihn als einen Fall der **vollständigen Korrelation**. Bei $r = +1$ wachsen beide Variable gleichzeitig im Durchschnitt (Fig. 23, α), bei $r = -1$ nimmt mit dem Wachsen der einen Variablen die andere ab (Fig. 23, β).

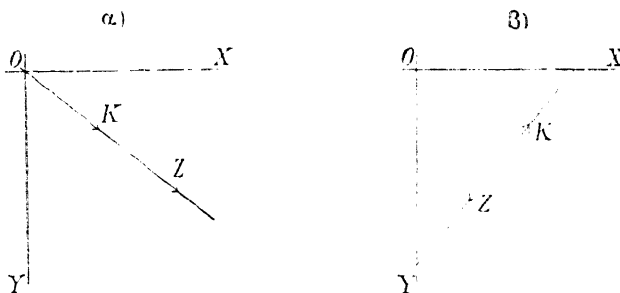



Fig. 23. Vollständige positive und negative Korrelation.

76. Durch die vorstehenden Feststellungen ist die Bedeutung von r und seine Bezeichnung als Korrelationskoeffizient gerechtfertigt. Der Definitionsgleichung (5) gemäß ist r eine absolute Zahl, weil Zähler und Nenner gleicher Dimension sind. Es ist daher gleichgültig, in welcher Einheit die Abweichungen ausgedrückt werden; nur muß dies bei x und μ_X, ebenso bei y und μ_Y einheitlich geschehen: dies ist bei der Berechnung von r wohl zu beachten.

Die Bedeutung der Größen b_1, b_2 ist die folgende. Die erste gibt ein Maß für die Abweichung der einzelnen Zeilenmittel vom allgemeinen Mittel M_X, die zweite ein Maß für die Abweichung der Kolonnenmittel vom allgemeinen Mittel M_Y. Wenigstens eine von diesen beiden Größen muß ein echter Bruch sein; dies geht aus ihren Ausdrücken (6) hervor, in denen r einen echten Bruch bedeutet und einer der beiden zueinander reziproken Quotienten $\frac{\mu_X}{\mu_Y}$, $\frac{\mu_Y}{\mu_X}$ unter 1 liegen muß; die andere Größe kann größer oder kleiner als 1 oder gleich 1 sein. Beide sind schließlich gleich bezeichnet und haben das Vorzeichen von r. Unbenannte Zahlen sind sie nur dann, wenn die Variablen X, Y gleichartige Größen sind.

Was die weitere Namengebung anlangt, so hat ein besonderer Fall, die Untersuchung der Erblichkeit der Körpergröße, Galton veranlaßt, die Größen b_1, b_2 als Regressionskoeffizienten, die Geraden ZZ', KK' als Regressionsgeraden zu bezeichnen, von einer Vorstellung ausgehend, die sich später als unhaltbar erwiesen hat: der Vorstellung einer Art Rückgangs der Nachkommenschaft gegenüber den Eltern. Wäre nämlich X die Körpergröße des Vaters, Y die Körpergröße des Sohnes, x die Abweichung der ersteren von dem allgemeinen Mittel M_X, so gibt die zweite der Gleichungen an, um wieviel dabei die durchschnittliche Größe der Söhne von deren allgemeinem Mittel M_Y differiert, nämlich um $y = b_2 x$; diese Abweichung nannte Galton die Regression. In analoger Weise wäre in dem besprochenen Falle $x = b_1 y$ die Regression in der Körpergröße des Vaters gegenüber der Abweichung y in der Größe des Sohnes. Yule schlägt für ZZ', KK' den Namen „charakteristische Linien“ und dementsprechend für die Gleichungen (7) den Namen „charakteristische Gleichungen“ vor [1]. Die Regressionsgerade entspricht der Trendgeraden [2], die in der modernen Wirtschaftsforschung eine Rolle spielt.

Die Hauptfunktion dieser Gleichungen besteht nach dem Gesagten darin, zu einer gegebenen Änderung der einen Variablen die durchschnittliche Änderung der andern abzuschätzen, und die Genauigkeit der Schätzung ist gleichbedeutend mit der Genauigkeit in der Bestimmung der Geraden, für welche die Ausdrücke (8) und (9) als maßgebend erkannt worden sind; dividiert man diese durch N und zieht die Quadratwurzel, so erhält man nach Art von mittleren Abweichungen die Größen

$$m_X = \mu_X \sqrt{1 - r^2} \qquad m_Y = \mu_Y \sqrt{1 - r^2} \tag{11}$$

als taugliche Maße zur Schätzung dieser Genauigkeit.

[1]) G. U. Yule, An Introduction to the Theory of Statistics. London 1932, p. 177.

[2]) Vgl. hierzu A. Timpe, Einführung in die Finanz- und Wirtschaftsmathematik. Berlin 1934, S. 182 u. f.

Eine andere Auffassung hat W. Wirth der Korrelation gegeben in seiner bemerkenswerten Schrift „Spezielle psychophysische Maßmethoden".[1] In dem Gebiet, das er mit dieser Schrift der mathematischen Behandlung auf statistischer Grundlage erschlossen hat, spielt die Frage nach dem Verhältnis der Änderungen zweier in Korrelation stehenden psychophysischen Größen eine wesentliche Rolle. Diese Frage kommt darauf zurück, den Beobachtungspunkten, welche durch die Wertepaare der beiden Variablen dargestellt sind, eine Gerade anzupassen: der Richtungskoeffizient dieser Geraden stellt dann das gesuchte Änderungsverhältnis dar.

Die vorgeführte Theorie der Regressionslinien könnte als eine Behandlung dieser Aufgabe angesehen werden, aber als eine solche, die aus mancherlei Gründen nicht befriedigt.

1. Führt sie im allgemeinen zu zwei verschiedenen Lösungen;

2. behandelt sie die beiden Variablen in ungleichartiger Weise, indem sie jedesmal nur die eine als „beobachtet" ansieht und ihre Abweichungen ausgleicht, während die andere als feststehend gilt;

3. ihre Ergebnisse sind vom Koordinatensystem abhängig, würden also andere werden, wenn man zu einem andern rechtwinkligen Koordinatensystem überginge.

Diese Umstände widerstreben der Natur des gestellten Problems, das nach einer einheitlichen Lösung verlangt.

Eine solche gibt ihm Wirth in der Weise, daß er nach jener Geraden sucht, der die Beobachtungspunkte „möglichst nahe liegen" in dem Sinne, daß die Quadratsumme ihrer Entfernungen von ihr zu einem Minimum wird. Er bezeichnet die so gefundene Linie als die „mittlere Gerade", ihren Richtungskoeffizienten als das „mittlere Änderungsverhältnis" der beiden in Beziehung gesetzten Variablen, und es zeigt sich, daß diese Gerade stets zwischen die beiden Regressionslinien zu liegen kommt.

In den Fällen, wo die Frage nach einem mittleren Verhältnis der Änderungen Sinn und Berechtigung hat, zeigen die Beobachtungspunkte die Tendenz, sich um eine Gerade zusammenzudrängen; dann weisen auch die Regressionslinien nahe übereinstimmende Richtungen auf und führen zu zwei nur wenig verschiedenen Proportionalitätsfaktoren. Im Sinne dieser Auffassung erblickt Wirth in dem Korrelationskoeffizienten ein Maß der Zweideutigkeit: je größer der Korrelationskoeffizient, desto weniger weichen die Regressionslinien voneinander ab, desto geringer ist die Zweideutigkeit, die in ihnen gelegen ist. Im Grenzzustande, d. i. wenn der Korrelationskoeffizient den Betrag 1 erreicht, verschwindet die Zweideutigkeit, die Regressionslinien vereinigen sich miteinander und mit der mittleren Geraden zu einer Linie, zwischen den beiden Variablen herrscht nicht bloß Korrelation, sondern funktionale, und zwar lineare, Abhängigkeit.[2]

---

[1]) Das Werk bildet einen Teil von E. Abderhaldens „Handbuch der biologischen Arbeitsmethoden", Berlin-Wien, 1920.

[2]) Auf die Beziehung, in welcher diese Wirthsche Auffassung zu der „normalen Korrelation" steht, werden wir in Art. 132 noch zurückkommen. Für ein näheres Eingehen auf die hier berührten Fragen mit Einschluß der Genauigkeitsbestimmungen für den Korrelationskoeffizienten, die Richtungsgrößen der Regressionslinien und der mittleren Geraden sei vor allem auf Wirths Werk selbst, insbesondere § 16, c), auf Czubers Ausführungen „Zur Theorie der linearen Korrelation" im Archiv für Psychologie, Bd. XLI, 1921, und auf Wirths Bemerkungen dazu an gleicher Stelle hingewiesen.

§ 6. Korrelation zwischen zwei Variablen.

Praktische Durchführung.

77. Bei der Anwendung der vorgeführten Theorie auf einen praktischen Fall wird es von der Fragestellung abhängen, wie weit man zu gehen hat, aber auch von dem Umfang der Erhebungen, wie weit man gehen kann.

Will man bloß feststellen, ob zwischen zwei Variabeln Korrelation besteht und welcher Art sie ist, so kann dies bei Kollektiven geringen Umfangs in der Weise geschehen, daß man die Wertepaare nach der einen Variablen, z. B. X, steigend ordnet; weisen dabei die Werte der andern Variablen, hier Y, deutlich steigende Tendenz auf, so kann man feststellen, daß positive Korrelation besteht; lassen sie deutlich fallende Tendenz erkennen, so kann negative Korrelation konstatiert werden. Ist weder das eine noch das andere mit Sicherheit auszusagen, so kann ohne Rechnung nichts Entscheidendes behauptet werden.

Aus den Messungen Charliers an Trientalis europaea[1]) seien zwei Reihen in dieser Form vorgeführt.

An 11 Exemplaren wurde Länge (X) und Dicke (Y) des Wurzelstockes gemessen; die Ergebnisse waren:

X (in mm)	Y (in mm)
3	2,0
5	2,3
5	2,2
6	2,0
7	1,4
7	1,6
7	3,0
8,5	1,9
9	2,0
10	3,0
10	2,6:

es ist keine Korrelation erkennbar.

An 13 Blüten ist die Länge des Stempels (X) und die Länge der Staubfäden (Y) gemessen worden mit folgenden Ergebnissen:

X (in mm)	Y (in mm)	
3,3	4,0	
4,0	4,4	15,3
4,1	3,5	
4,2	3,4	

[1]) C. V. L. Charlier, A Statistical Description of Trientalis europaea, Arkiv för Botanik, Bd. XII, Nr. 14, 1913, S. 7 und 21.

X (in mm)	Y (in mm)	
4,2	4,3	18,8
4,2	4,8	
4,3	5,1	
4,4	4,6	
4,6	5,4	19,8
4,9	3,8	
5,0	5,1	
5,0	5,5	
5,2	5,2;	

trotz der Schwankungen in der Reihe Y kann man deutlich eine steigende Tendenz wahrnehmen, besser als in den Einzelwerten in den Gruppensummen; man kann also sagen, daß zwischen den beiden Charakteren positive Korrelation besteht.

Will man sich mit einer beiläufigen Bestimmung von r begnügen, so kann das folgende geometrische Verfahren eingeschlagen werden. Man bestimmt die Reihenmittel $M_X^{(Y)}$, $M_Y^{(X)}$ und die Summenmittel M_X, M_Y, trägt alle im Koordinatensystem ein und sucht mittels eines über M ausgespannten dünnen schwarzen Fadens den Punktreihen der Zeilen- und Kolonnenmittel so nahe zu kommen als möglich, indem man den Faden jedesmal so lange um M dreht, bis sich die beiderseitigen Abweichungen, so gut man es beurteilen kann, ausgleichen. Man zieht die so bestimmten Geraden und mißt die Winkel α_1, α_2, Fig. 21: sodann nimmt man ihre Tangenten b_1, b_2 und hat gemäß den Formeln (6)

$$b_1\, b_2 = r^2,$$

also $|r| = \sqrt{b_1\, b_2}$; das Vorzeichen von r ist $+$, wenn tg α_1, tg α_2 positiv sind (wie in der Figur), hingegen $-$, wenn beide negativ sind.

78. Die vollständige rechnerische Durchführung eines Falles hat mit der Bestimmung von M_X, M_Y, μ_X, μ_Y zu beginnen; durch das erste Größenpaar ist der Mittelpunkt M der Tafel, der Ausgangspunkt der weiteren Rechnungen und der Träger der Konstruktion, gefunden. Die mühevollste Arbeit bei ausgedehnten Tafeln ist die Ausrechnung des Produktmittels p, eine Angelegenheit für sich, mit der wir uns nun befassen wollen.

Der direkte Weg wäre der folgende: Man bestimmt zu jedem besetzten Punkt die relativen Koordinaten (Abweichungen) in Bezug auf M, $x = X - M_X$, $y = Y - M_Y$; bildet ihr Produkt unter sorgfältiger Beachtung der Vorzeichen, multipliziert es mit dem zugehörigen z, bildet die algebraische Summe all dieser Produkte und dividiert sie durch N; damit hat man p erhalten.

So angelegt aber würde sich die Rechnung in unbequemen Zahlen bewegen und viel Mühe und angespannte Aufmerksamkeit erfordern.

Eine wesentliche Vereinfachung wird dadurch erzielt, daß man nicht M, sondern den Mittelpunkt $\mathfrak{M}$ jenes Feldes zum Ausgangspunkt nimmt, in welchem M liegt, und daß man nicht in den natürlichen Einheiten, sondern in Klassengrößen rechnet; das hat den großen Vorteil, daß sich nunmehr die Abweichungen, das sind die relativen Koordinaten der besetzten Punkte in Bezug auf $\mathfrak{M}$, die $\mathfrak{x}$, $\mathfrak{y}$ heißen mögen, in ganzen Zahlen ausdrücken, deren Produkte sich unmittel-

bar zu den Häufigkeitszahlen hinschreiben lassen[1]). Damit ist das Material zur Bildung jener Summe gegeben, die in dem Ausdruck für p auftritt; nur muß die Summe auf den wahren Ausgangspunkt M zurückgeführt werden, bevor man weiter rechnet. Dies geschieht in der folgenden einfachen Weise.

Gesucht soll werden $p = \frac{1}{N} \Sigma (z\,x\,y)$, gerechnet aber wird $\mathfrak{p} = \frac{1}{N} \Sigma (z\,\mathfrak{x}\,\mathfrak{y})$; nennt man die relativen Koordinaten von $\mathfrak{M}$ in Bezug auf M: $\mathfrak{a}$, $\mathfrak{b}$, so ist

$$\mathfrak{x} = x - \mathfrak{a}$$
$$\mathfrak{y} = y - \mathfrak{b},$$

daher

$$\mathfrak{x}\mathfrak{y} = xy - \mathfrak{b}x - \mathfrak{a}y + \mathfrak{a}\mathfrak{b}$$

und

$$\Sigma (z\,\mathfrak{x}\,\mathfrak{y}) = \Sigma (z\,x\,y) - \mathfrak{b}\,\Sigma (z\,x) - \mathfrak{a}\,\Sigma (z\,y) + \mathfrak{a}\mathfrak{b}\,\Sigma (z);$$

nun ist aber $\Sigma (z\,x) = 0$ und $\Sigma (z\,y) = 0$, weil M der Mittelpunkt, daher weiter

$$\Sigma (z\,\mathfrak{x}\,\mathfrak{y}) = \Sigma (z\,x\,y) + N\mathfrak{a}\mathfrak{b};$$

Division durch N gibt

$$\mathfrak{p} = p + \mathfrak{a}\mathfrak{b},$$

daher ist schließlich

$$p = \mathfrak{p} - \mathfrak{a}\mathfrak{b}. \qquad (12)$$

Man hat also von dem bezüglich des willkürlichen Ausgangspunktes $\mathfrak{M}$ gebildeten Produktmittel $\mathfrak{p}$ das Produkt der relativen Koordinaten von $\mathfrak{M}$ in Bezug auf M unter genauer Beachtung ihrer Vorzeichen zu subtrahieren, um das gesuchte p zu erhalten.

79. Die Anlage der Rechnung kann am besten in Verbindung mit praktischen Fällen gezeigt werden. Das soll in einer Reihe von Beispielen geschehen, die zugleich eine Vorstellung von verschiedenen Graden und den beiden Arten der Korrelation geben werden.

Beispiele. 1) Beginnen wir mit der kleinen Korrelationstafel in Art. 72, Tab. 52), betreffend die Stammdicke X und Länge des längsten Blumenblattes Y bei Trientalis europaea.

Die Rechnung ergibt die folgenden Fundamentalgrößen:

$$M_X = 0{,}813 \text{ mm}, \qquad \mu_X = 1{,}894 \text{ Klassen} = 0{,}189 \text{ mm}$$
$$M_Y = 36{,}244 \text{ mm}, \qquad \mu_Y = 1{,}823 \quad \text{„} \quad = 10{,}94 \text{ mm}:$$

man wird daher als Ausgangspunkt $\mathfrak{M}$ die Feldmitte 0,825/34,5 wählen und durch das in starken Linien gezeichnete Kreuz die Reihen abgrenzen, die durch diesen Punkt gehen. Die Häufigkeitszahlen sind in den einzelnen Feldern mit klein gedruckten Ziffern eingetragen, die Abweichungsprodukte in fetten Ziffern dazugesetzt. Ihr Fortschreiten wird sofort klar, wenn man sich zu jedem Felde die Abweichungen $\mathfrak{x}$, $\mathfrak{y}$ hinzudenkt. Innerhalb des Kreuzes sind alle Abweichungsprodukte Null, weil die eine oder die andere Abweichung oder beide zugleich

[1]) Bei einem unstetigen Kollektiv wird $\mathfrak{M}$ nach dem Gitterpunkt verlegt, der M am nächsten liegt.

(im Kreuzungsfeld) verschwinden. Auf die Vorzeichen der Produkte ist keine Rücksicht genommen: sie sind in dem rechten untern Quadranten (I) und in dem diagonal gegenüberliegenden (III) durchwegs +, in dem linken untern (II) und dem diagonal gegenüberliegenden (IV) durchwegs —. In der Reihenfolge dieser Quadranten I—IV werden die Produkte mit den Häufigkeitszahlen gebildet und erst deren Summen mit dem zugehörigen Vorzeichen versehen, wie dies unter der Tabelle geschehen ist.

Tab. 52a. Bestimmung des Korrelationskoeffizienten.

Länge des längsten Blumenblattes (in mm)	Stammdicke (in mm)											Summe
	0,425	0,525	0,625	0,725	0,825	0·925	1·025	1·125	1·225	1·325	1·425	
10.5	1 16	.	.	.	.	.	.	.	.	.	.	1
16,5	1 12	4 9	1 6	1 3	.	.	.	.	.	.	.	7
22,5	1 8	9 6	16 4	8 2	1 0	.	.	.	.	.	.	30
28,5	.	2 3	9 2	22 1	9 0	2 1	1 2	.	.	.	.	45
34.5	.	.	8 0	19 0	20 0	4 0	1 0	.	.	.	.	52
40.5	1 4	.	.	7 1	18 0	12 1	6 2	4 3	.	.	.	48
46,5	.	.	.	1 2	8 0	9 2	3 4	2 6	1 8	.	.	24
52,5	.	.	.	.	.	3 3	6 6	4 9	1 12	.	.	14
58,5	.	.	.	.	.	.	2 8	2 12	1 16	2 20	.	7
64,5	.	.	.	.	.	.	.	.	1 20	3 25	.	4
70,5	.	.	.	.	.	.	.	.	1 24	.	1 36	2
Summe	4	15	34	53	56	30	19	12	5	5	1	234

I		III		II	IV
12	12	22	8	7	2
12	16	18	3	4	2
12	24	6	6	2	4
18	16	6	36	13	
12	40	64	12		
12	20	54	16		
8	75		+251		
9	24				
36	36				
36	+430				

I + 430
III + 251
+ 681
II — 13
IV — 4
+ 664

664 : 234 = 2,8376

Hiernach ist

$$\mathfrak{p} = 2{,}8376,$$

ferner ist

$$\mathfrak{a} = \frac{0{,}825 - 0{,}813}{0{,}1} = + 0{,}12 \qquad \mathfrak{b} = \frac{34{,}5 - 36{,}244}{6} = - 0{,}2907 \text{ in Klassen}$$

$$\mathfrak{a}\,\mathfrak{b} = - 0{,}0349,$$

somit

$$p = 2{,}8376 + 0{,}0349 = 2{,}8725$$

und schließlich der Korrelationskoeffizient

$$r = \frac{2{,}8725}{1{,}894 \cdot 1{,}823} = 0{,}832.$$

Es besteht also zwischen den beiden in Beziehung gesetzten Merkmalen eine hohe positive Korrelation, wie dies schon das äußere Bild der Tabelle erwarten ließ, auf das gleich bei der ersten Vorführung hingewiesen wurde.

2) Zu der Tab. 53, dieselbe Pflanze betreffend, seien nur die Hauptergebnisse der Rechnung hierhergesetzt:

$$M_X = 14{,}03 \text{ mm}, \quad \mu_X = 2{,}191 \text{ Klassen} = 4{,}38 \text{ mm}$$
$$M_Y = 35{,}53 \text{ mm}, \quad \mu_Y = 2{,}222 \text{ Klassen} = 11{,}11 \text{ mm}$$
$$p = 4{,}3403$$
$$r = 0{,}892.$$

Die letzte Zahl zeigt, daß zwischen Breite und Länge des längsten Blumenblattes nahezu derselbe Grad von Korrelation besteht, wie er im vorigen Beispiel gefunden wurde.

3) Zu einem weiter ausgeführten Beispiel sei die Tab. 51 benützt, betreffend eine etwa bestehende Abhängigkeit zwischen der Fruchtbarkeit der Väter und ihrer Söhne. Hier entfallen alle Reduktionen wegen der Klassengröße, weil diese bei beiden Variablen mit 1 anzusetzen ist.

Die Grundgrößen sind schon in Art. 74 angegeben worden, u. zw.

$$M_X = 5{,}83 \qquad M_Y = 5{,}07$$
$$\mu_X = 2{,}90 \qquad \mu_Y = 3{,}15;$$

demzufolge wird zum Ausgangspunkt $\mathfrak{M}$ für die Berechnung von r der Gitterpunkt 6/5 zu nehmen sein mit den relativen Koordinaten $\mathfrak{a} = 0{,}17$, $\mathfrak{b} = - 0{,}07$.

Die zur Berechnung des Produktmittels p führende Tabelle gestaltet sich demnach wie folgt.

Tab. 51a. Bestimmung des Korrelationskoeffizienten.

Zahl der Kinder des Sohnes: Y	Zahl der Kinder des Vaters: X																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	5 **25**	8 **20**	7 **15**	14 **10**	18 **5**	2 **0**	2 **5**	3 **10**	8 **15**	3 **20**	4 **25**	4 **30**	.	.	.	.	.
1	3 **20**	8 **16**	6 **12**	5 **8**	8 **4**	8 **0**	6 **4**	5 **8**	4 **12**	.	2 **20**	.	1 **28**	.	.	.	.
2	7 **15**	5 **12**	6 **9**	13 **6**	12 **3**	12 **0**	12 **3**	6 **6**	5 **9**	4 **12**	2 **15**	1 **18**	1 **21**	.	.	.	.
3	5 **10**	10 **8**	13 **6**	11 **4**	17 **2**	13 **0**	13 **2**	12 **4**	10 **6**	4 **8**	1 **10**	2 **12**	1 **14**	.	.	.	.
4	4 **5**	16 **4**	18 **3**	24 **2**	23 **1**	5 **0**	18 **1**	10 **2**	7 **3**	8 **4**	1 **5**	5 **6**	2 **7**	1 **8**	.	.	1 **11**
5	9 **0**	8 **0**	11 **0**	14 **0**	16 **0**	12 **0**	16 **0**	12 **0**	2 **0**	5 **0**	8 **0**	6 **0**	3 **0**	.	2 **0**	.	.
6	3 **5**	4 **4**	10 **3**	16 **2**	13 **1**	11 **0**	11 **1**	10 **2**	10 **3**	1 **4**	2 **5**	2 **6**	1 **7**	.	.	.	.
7	5 **10**	6 **8**	8 **6**	7 **4**	10 **2**	14 **0**	11 **2**	12 **4**	4 **6**	7 **8**	1 **10**	1 **12**		.	.	.	.
8	3 **15**	5 **12**	4 **9**	15 **6**	19 **3**	7 **0**	10 **3**	8 **6**	4 **9**	2 **12**	2 **15**	1 **18**	.	.	1 **27**	.	.
9	1 **20**	6 **16**	5 **12**	9 **8**	5 **4**	5 **0**	8 **4**	5 **8**	4 **12**	3 **16**	3 **20**	1 **24**	.	.	.	.	.
10	2 **25**	3 **20**	9 **15**	3 **10**	2 **5**	5 **0**	4 **5**	6 **10**	4 **15**	3 **20**	1 **25**	1 **30**	1 **35**	.	.	.	.
11	1 **30**	1 **24**	1 **18**	4 **12**	2 **6**	1 **0**	1 **6**	2 **12**	1 **18**	1 **24**	.	.	.	.	.	.	.
12	.	.	2 **21**	2 **14**	2 **7**	.	1 **7**	1 **14**	1 **21**	2 **28**	.	1 **42**	.		.		
13	1 **40**	.	.	1 **16**	3 **8**	.	1 **8**	1 **16**	1 **24**	1 **32**	.	1 **48**		1 **64**	.	.	.
14	.	1 **36**	.	.	.	1 **0**	.	.	.	.	1 **45**	.	.	.	.	.	.
15	.	.	1 **30**	.	.	.	.	1 **20**	.	.	1 **50**	.		.			.
16	.	.	.	.	.	.		.	1 **33**	1 **44**		.	.	.	.	.	.

Daraus fließen die folgenden Produkte:

I				II			III			IV	
11	14	24	12	13	28	60	23	105	18	120	14
22	16	48	12	20	16	24	34	64	26	32	14
30	20	60	18	57	30	36	36	80	36	32	21
32	30	24	24	20	48	15	32	60	24	48	28
20	24	56	30	10	36	50	90	48	10	60	8
6	36	32	42	12	60	45	48	160	20	5	11
7	48	44	48	14	135	20	44	20	48	10	1227
8	60	10	7	24	18	50	78	50	36	30	—
20	18	10	35	32	42	30	40	105	40	40	
48	21	30	64	28	30	40	140	60	30	100	
48	24	60	27	90	16	1503	54	125	21	30	
40	33	25	1617	72	48	-	78	1700	60	24	
60	4	45	+	30	60		54	+	45	18	
24	56	50		48	96		72		48	120	

Die weitere Rechnung verläuft so:

$$\mathfrak{p} = 0{,}587$$

$$\mathfrak{a}\,\mathfrak{b} = -0{,}012$$

$$p = 0{,}599$$

$$r = 0{,}0656$$

$$b_1 = 0{,}0604 \qquad b_2 = 0{,}0713$$

$$\operatorname{tg}\Theta = \frac{0{,}9957}{2{,}0068 \cdot 0{,}0656}$$

$$\Theta = 82^{0}\,28'.$$

Die charakteristischen Linien haben die Gleichungen

$$\xi = 0{,}0604\,\eta \qquad \eta = 0{,}0713\,\xi$$

und sind unter den Winkeln $3^0\,27'$, beziehungsweise $4^0\,5'$ zur Vertikalen, beziehungsweise zur Horizontalen geneigt. Sie besagen: Nimmt die Zahl der Kinder des Vaters um 1 zu, so entspricht dem ein Zuwachs der durchschnittlichen Zahl der Kinder des Sohnes um 0,0713, und nimmt die Zahl der Kinder des Sohnes um 1 zu, so ist der durchschnittliche Zuwachs in der Kinderzahl des Vaters 0,0604.

Es besteht hiernach zwischen der Fruchtbarkeit der Väter und ihrer Söhne nur eine schwache positive Korrelation. Fig. 24 zeigt den Sachverhalt im geometrischen Bilde. Die geringelten Punkte bezeichen die Kolonnenmittel, die mit Kreuzchen versehenen die Zeilenmittel. Die Maße der Annäherung der Geraden K, Z an diese Punktfolgen:

$$m_Y = 3{,}14$$
$$m_X = 2{,}89$$

bestätigen den Eindruck des Bildes, daß sich die Gerade Z der sie betreffenden Punktreihe etwas besser anpaßt als K der ihrseitigen.

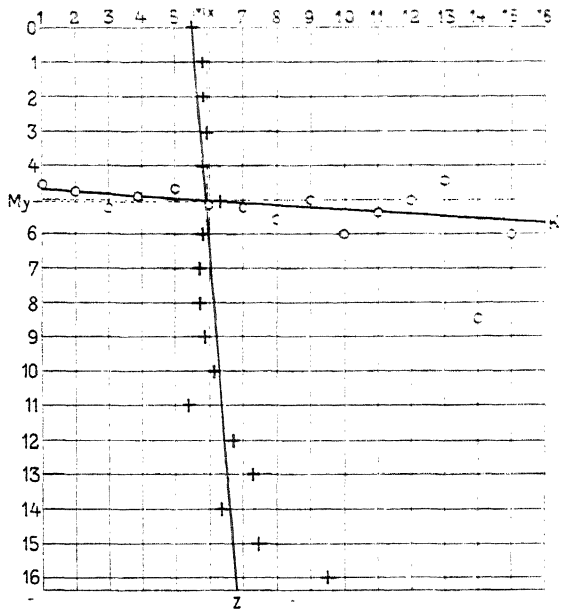



Fig. 24. Geometrisches Bild zur Tab. 51a.

4) Die gleiche Durcharbeitung der Tab. 55, Art. 74, über die Fruchtbarkeit der Mütter und ihrer Töchter führt zu folgenden Resultaten, die nach dem vorstehenden Beispiel keiner näheren Erklärung bedürfen.

$$M_X = 5{,}90 \qquad M_Y = 4{,}34$$
$$\mu_X = 2{,}83 \qquad \mu_Y = 2{,}97$$

$\mathfrak{M}(6/4)$: $\mathfrak{a} =$ 0,10 $\mathfrak{b} =$ ·

$\mathfrak{p} =$ 1,754

$\mathfrak{ab} =$ · 0,034

$p =$ 1,788

$r =$ 0,2127

$$b_1 = 0{,}2027 \qquad b_2 = 0{,}2232$$

$$\operatorname{tg}\Theta = \frac{0{,}9548}{0{,}2027 + 0{,}2232}$$

$$\Theta = 65^0\, 58'$$

$$\xi = 0{,}2027\,\eta \qquad \eta = 0{,}2232\,\xi$$
$$m_X = 2{,}76 \qquad m_Y = 2{,}90.$$

Die Vergleichung dieser Ergebnisse mit den vorigen führt zu dem Schlusse, daß in der weiblichen Fruchtbarkeit eine schärfer und stärker ausgeprägte posi-

tive Korrelation besteht. Man vergleiche dazu das geometrische Bild Fig. 25; hier besteht in der Verläßlichkeit der Bestimmung der Geraden K, Z nur ein unwesentlicher Unterschied.

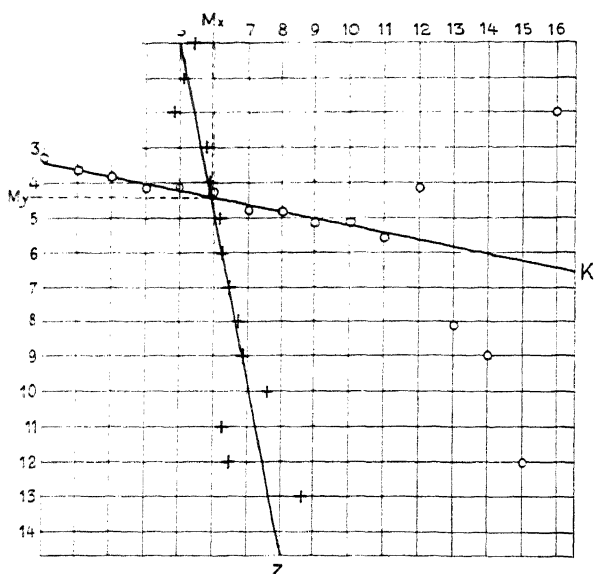



Fig. 25. Geometrisches Bild zur Tab. 55.

5) Bei der Bearbeitung der Tab. 54, Art. 72, 4), welche die Blätter des wilden Efeu zum Gegenstande hat, ist wieder auf die Klassengröße zu achten. Die Durchführung der Arbeit dem Leser überlassend, seien wieder nur die wesentlichen Zwischen- und Endergebnisse hier zusammengestellt.

$M_X = 10{,}90$, $\mu_X = 1{,}71$ Klassen $= 3{,}42$ Achtelzoll

$M_Y = 13{,}23$, $\mu_Y = 2{,}28$ „ $= 4{,}56$ „

$\mathfrak{M}$ (9,95/13,95), $a = -\ 0{,}475$ $b = 0{,}36$ in Klassen

$$\mathfrak{p} = 3{,}1724$$
$$ab = -0{,}1710$$
$$p = 3{,}3434$$
$$r = 0{,}8575$$

$b_1 = 0{,}6431$ $b_2 = 1{,}1433$

$$\operatorname{tg}\Theta = \frac{0{,}2647}{1{,}7864}$$

$$\Theta = 8^{\circ}26'$$

$\xi = 0{,}6431\,\eta$ $\eta = 1{,}1433\,\xi$

$m_X = 0{,}8796$ $m_Y = 1{,}1728.$

Der hohe Grad positiver Korrelation zeugt für das Streben der Natur, das Verhältnis zwischen Länge und Breite der Efeublätter möglichst konstant zu erhalten. Die diesem Fall entsprechende Fig. 26 zeigt ein ganz anderes Bild als die früheren.

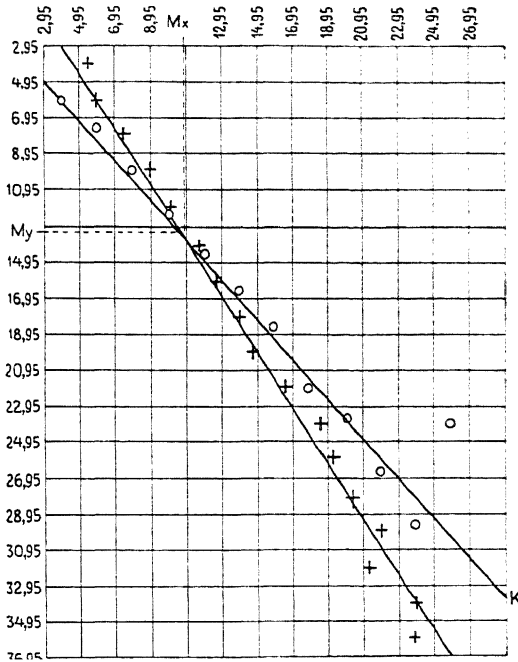



Fig. 26. Geometrisches Bild zur Tab. 54.

80. In Art. 75 ist bereits darauf hingewiesen worden, daß in der Annahme, die Reihenmittel würden auf geraden Linien liegen, eine Willkür enthalten ist, der freilich zahlreiche Erfahrungen gegenüberstehen, die diese Annahme des Vorhandenseins einer linearen Korrelation gerechtfertigt erscheinen lassen. Es sind aber auch schon Fälle bekannt geworden, wo die eine oder die andere oder auch beide Reihen von Mitteln auf krumme Linien mehr oder weniger deutlich hinweisen. Es steht nichts im Wege, auch dann die vorgetragene Theorie zur Anwendung zu bringen; doch darf dies bei der Wertung der Resultate nicht übersehen werden: insbesondere geht dann die Vergleichbarkeit mit einer linearen Korrelation verloren. Daß solche besondere Fälle ein erhöhtes Interesse beanspruchen, liegt auf der Hand.

Vorerst soll eine Materie vorgeführt werden, welche diese Erscheinung in einem mäßigen Grade aufweist. In der Gebäranstalt in Lund sind die Gewichte neugeborener Knaben mit den zugehörigen Plazentagewichten aufgezeichnet worden, um zu erforschen, ob zwischen den beiden Gewichten eine Korrelation stattfindet. Die folgende Tabelle zeigt die Verteilung der 1223 Fälle und gibt die Reihenmittel an[1].

[1] S. D. Wicksell, Meddelande från Lunds Astronomiska Observatorium, Nr. 80, 1917.

Wenn schon der bloße Anblick dieser Tabelle, aus der unzweifelhaft eine positive Korrelation hervorgeht, eine gewisse „Krümmung“ in der Ziffernmasse erkennen läßt, so wird dies noch deutlicher aus der geometrischen Darstellung, die in Fig. 27 gegeben ist. Beide Linien, Z und K, weisen eine ausgesprochene Krümmung auf und ihre Form würde bei einer Vermehrung des Materials bestimmter hervortreten. Die gezogenen Netzlinien sind Mittellinien der Felder.

Tab. 56. Korrelation zwischen dem Gewicht neugeborener Knaben und dem Gewicht der Plazenta.

Gewicht der Plazenta (in g): Y	Gewicht des Kindes (in g): X											Summe	$M_X^{(Y)}$
	1850	2150	2450	2750	3050	3350	3650	3950	4250	4550	4850		
300	1	.	2	.	1	.	.	.	.	.	.	4	2450
380	2	10	8	25	21	14	2	.	.	.	.	82	2827
460	2	2	12	34	94	77	37	10	2	.	.	270	3180
540	.	1	4	17	55	111	84	37	10	.	1	320	3433
620	.	1	3	6	24	51	92	78	22	3	.	280	3647
700	.	2	1	2	4	15	40	51	24	8	4	151	3841
780	.	.	.	.	2	8	13	26	20	11	3	83	4008
860	.	.	.	.	.	4	1	8	3	6	4	26	4158
940	.	.	.	.	.	1	1	3	.	.	2	7	4079
Summe	5	16	30	84	201	281	270	213	81	28	14	1223	
$M_Y^{(X)}$	396	455	463	470	500	546	593	651	678	757	786		

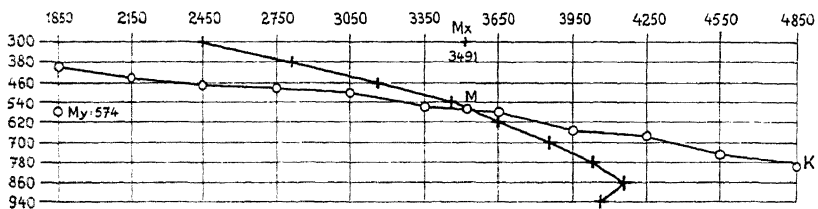



Fig. 27. Gewichte neugeborener Knaben und der Plazenta.

Eine weit auffälligere Besonderheit äußert sich in dem folgenden Beispiel, das in allen Einzelheiten durchgearbeitet werden soll.

Für 324 sächsische Gemeinden mit einer Einwohnerzahl von 2000 bis 50000 ist für das Jahr 1935 berechnet worden, wieviel Knaben sich unter 100 Geborenen befanden [1]. Untersucht soll werden, ob diese Verhältniszahlen in Korrelation stehen mit den Gesamtgeburtenzahlen der Gemeinden. Eine Häufigkeitszahl der folgenden Korrelationstabelle zeigt also an, in wieviel Gemeinden eine bestimmte Gesamtzahl von Geburten mit einem bestimmten Prozentsatz von Knaben zusammentraf. Bei der Berechnung des Prozentsatzes ist nicht aufgerundet worden, damit die Gruppenbildung sachgemäß erfolgen konnte; es wurde durchgängig die untere Gruppen-

[1] Zeitschrift des Sächsischen Statistischen Landesamtes, 82. Jahrg. 1936, S. 19.

grenze zur Gruppe gerechnet, die obere Gruppengrenze dagegen nicht. Zum Unterschiede von den bisherigen Beispielen sind im gegenwärtigen die beiden Variablen Größen verschiedener Art.

Die Tabelle ist so ausgestattet, daß sie alle für die weitere Rechnung erforderlichen Daten enthält, zu welchen noch die arithmetischen Mittel und die mittleren Abweichungen der Summenreihen kommen, nämlich:

$M_X = 51{,}48 \qquad \mu_X = 3{,}907$ Klassen $= 5{,}86$ Prozent

$M_Y = 109{,}72 \qquad \mu_Y = 4{,}381 \quad$ „ $\quad = 109{,}53$ Geburten

Fig. 28. Geometrisches Bild zur Tab. 57.

Tab. 57. Korrelation zwischen Geburtenmenge und Geschlechtsverhältnis.

Zahl der Geborenen: Y	Unter 100 Geborenen befinden sich Knaben: X																									Summe	$M_X^{(Y)}$	$\mu_X^{(Y)}$
	30,5–32,0	32,0–33,5	33,5–35,0	35,0–36,5	36,5–38,0	38,0–39,5	39,5–41,0	41,0–42,5	42,5–44,0	44,0–45,5	45,5–47,0	47,0–48,5	48,5–50,0	50,0–51,5	51,5–53,0	53,0–54,5	54,5–56,0	56,0–57,5	57,5–59,0	59,0–60,5	60,5–62,0	62,0–63,5	63,5–65,0	65,0–66,5	66,5–68,0			
0–25	.	.	.	1 10	.	.	.	.	1 20	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	39,500	3,750
25–50	1 39	1 36	.	.	.	4 24	1 21	6 18	9 15	.	6 9	7 6	3 3	7 0	1 3	8 6	4 9	7 12	6 15	7 18	3 21	.	4 27	2 30	3 33	90	52,067	7,365
50–75	.	.	.	.	1 18	3 16	2 14	3 12	3 10	2 8	5 6	6 4	4 2	6 0	8 2	11 4	5 6	8 8	3 10	1 12	4 14	1 16	2 18	.	.	78	51,385	6,360
75–100	.	.	.	.	.	.	.	1 6	.	3 4	2 3	9 2	3 1	7 0	5 1	4 2	5 3	.	1 5	1 6	2 7	1 8	.	.	.	44	51,295	4,650
100–125	.	.	.	.	.	.	.	.	2 0	3 0	4 0	3 0	5 0	1 0	2 0	.	2 0	5 0	1 0	2 0	1 0	.	.	.	.	31	51,089	5,355
125–150	.	.	.	.	.	.	.	.	.	1 4	.	5 2	3 1	3 0	3 1	.	1 3	2 4	2 5	.	.	.	.	.	.	20	51,275	3,825
150–175	.	.	.	.	.	.	.	.	.	.	.	1 4	1 2	2 0	3 2	2 4	2 6	1 8	.	.	.	.	.	.	.	12	52,500	2,513
175–200	.	.	.	.	.	.	.	.	.	.	.	.	3 3	3 0	1 3	.	.	1 12	.	.	.	.	.	.	.	8	51,125	2,340
200–225	.	.	.	.	.	.	.	.	.	.	.	.	3 4	.	.	1 8	1 12	1 16	.	.	.	.	.	.	.	6	52,250	3,120
225–250	.	.	.	.	.	.	.	.	.	1 20	1 15	1 10	.	1 0	.	1 10	1 15	.	.	.	.	.	.	.	.	6	49,750	3,840
250–275	.	.	.	.	.	.	.	.	.	.	.	1 12	1 6	1 0	.	.	1 18	1 24	.	.	.	.	.	.	.	5	51,950	3,480
275–300	.	.	.	.	.	.	.	.	.	.	.	.	1 7	1 0	1 7	1 14	.	1 28	.	.	.	.	.	.	.	5	52,550	2,580
300–325	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1 48	.	.	.	.	.	1	59,750	0
325–350	.	.	.	.	.	.	.	.	.	.	.	.	.	1 0	.	1 18	.	.	.	.	.	.	.	.	.	2	52,250	1,500

350—375	.	.	.	.	.	.	.	.	.	.	1 30	.	1 10	.	.	.	.	.	.	.	.	.	.	.	.	2	47,750	1,500
375—400	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
400—425	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1 12	.	.	.	.	.	.	.	.	.	.	1	52,250	0
425—450	.	.	.	.	.	.	.	.	.	.	.	1 26	.	.	1 13	.	.	.	.	.	.	.	.	.	.	2	50,000	2,250
450—475	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
475—500	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
500—525	.	.	.	.	.	.	.	.	.	1 64	.	.	.	.	.	1 32	.	.	.	.	.	.	.	.	.	2	49,250	4,500
525—550	.	.	.	.	.	.	.	.	.	.	.	1 34	.	.	.	.	.	.	.	.	.	.	.	.	.	1	47,750	0
550—575	.	.	.	.	.	.	.	.	.	.	.	.	1 18	.	.	.	.	.	.	.	.	.	.	.	.	1	49,250	0
575—600	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2 19	.	.	.	.	.	.	.	.	.	.	2	52,250	0
600—625	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1 20	.	.	.	.	.	.	.	.	.	.	1	52,250	0
625—650	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
650—675	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
675—700	.	.	.	.	.	.	.	.	.	.	.	.	.	1 0	1 23	.	.	.	.	.	.	.	.	.	.	2	51,500	0,750
Summe	1	1	.	1	1	7	3	10	15	11	19	35	29	34	30	30	22	27	13	12	10	2	6	2	3	324		
$M_Y^{(X)}$	37,50	37,50	.	12,50	62,50	48,21	54,17	50,00	50,83	146,59	92,76	116,79	153,02	131,62	196,67	108,33	104,55	100,46	68,27	79,17	65,00	75,00	45,83	37,50	37,50			
$\mu_Y^{(X)}$	0	0	.	0	0	12,38	11,78	16,78	27,18	124,50	79,00	105,50	110,03	123,75	190,50	107,50	64,25	68,75	36,88	76,00	23,58	12,50	11,78	0	0			

In Bezug auf $\mathfrak{M}$ (50,75/112,5) ergibt sich die Produktsumme

$$\mathfrak{p} = -0{,}2037$$

und wegen $a = \frac{50{,}75 - 51{,}48}{1{,}5} = -0{,}487 \qquad \mathfrak{b} = \frac{112{,}5 - 109{,}72}{25} = 0{,}111$

$$p = -0{,}1496,$$
$$r = -0{,}009.$$

$$b_1 = -0{,}009\frac{3{,}907}{4{,}381} = -0{,}0080, \quad b_2 = -0{,}009\frac{4{,}381}{3{,}907} = -0{,}0101$$

$$\operatorname{tg}\Theta = -\frac{1{,}0001}{0{,}0080 + 0{,}0101}$$

$$\Theta = 91^0\,2'$$

$$\xi = -0{,}0080\,\eta \qquad \eta = -0{,}0101\,\xi$$

$$m_X = 3{,}907 \qquad m_Y = 4{,}381$$

Der Fall ist in mehrfacher Beziehung von Interesse. Einmal als Bestätigung des Gesetzes der großen Zahlen: mit wachsender Gesamtzahl der Geburten wird das Intervall, innerhalb dessen das Geschlechtsverhältnis schwankt, im allgemeinen enger. Zweitens als Beispiel einer krummlinigen Anordnung der Reihenmittel wenigstens der einen Art, nämlich der Kolonnenmittel, während die Zeilenmittel sich einer Geraden anpassen; damit hängt auch der Unterschied (0,474) zwischen m_X und m_Y zusammen. Drittens: als Beispiel einer wenn auch sehr schwachen negativen Korrelation. Freilich kann angesichts der deutlichen krummlinigen Anordnung der einen Mittelreihe die Frage aufgeworfen werden, ob noch mit Berechtigung von dem Formelapparat Gebrauch gemacht werden kann, der sich auf die Voraussetzung angenähert geradliniger Anordnung stützt.

Die Fig. 28 bringt den Inhalt dieser eigenartigen Korrelationstabelle zu geometrischer Anschauung.

81. Im Anschlusse an das letzte Beispiel wollen wir uns mit einem Korrelationsmaße bekanntmachen, das frei ist von einer Annahme über die Anordnung der Reihenmittel, also auch unabhängig von der Hypothese ihrer geradlinigen Lagerung. Es ist das von Pearson eingeführte Korrelationsverhältnis. Hat man daneben auch den Korrelationskoeffizienten bestimmt, so gibt die Differenz beider Größen, und zwar der Überschuß der ersten über die zweite (wie sich weiter herausstellen wird) ein Maß der Abweichung der Reihenmittel von der geradlinigen Anordnung.

Fig. 29. Zur Ableitung des Korrelationsverhältnisses.

Richten wir unsere Aufmerksamkeit auf eine bestimmte Zeile (Y), auf das in ihr liegende Mittel $M_X^{(Y)}$, auf das allgemeine Mittel der X, nämlich M_X, und auf die Gerade Z, welche den Zeilenmitteln angepaßt worden ist, Fig. 29. Wir führen folgende Bezeichnungen ein:

x sei die Abweichung eines X in der Zeile vom allgemeinen Mittel M_X;
x' seine Abweichung vom Zeilenmittel $M_X^{(Z)}$;
x'' seine Abweichung von der Geraden Z;
d die Abweichung des Zeilenmittels vom allgemeinen Mittel;
δ die Abweichung des Zeilenmittels von der Geraden Z;
n die Häufigkeitszahl der Zeile;
N der Umfang der Tafel.

Dann ist

$$\mu_X^{(Z)^2} = \frac{1}{n} \Sigma (x'^2) \tag{13}$$

das mittlere Abweichungsquadrat in der Zeile und

$$\mu_X^2 = \frac{1}{N} \Sigma (x^2) \tag{14}$$

das allgemeine mittlere Abweichungsquadrat, Größen, die früher schon bestimmt und verwendet wurden.

Bilden wir den Mittelwert der $\mu_X^{(Z)^2}$ unter Berücksichtigung der Zeilenhäufigkeit als Gewicht und nennen ihn $\mu_X^{(\mathfrak{z})^2}$, so ist

$$\mu_X^{(\mathfrak{z})^2} = \frac{1}{N} \Sigma \left(n\, \mu_X^{(Z)^2}\right), \tag{15}$$

die Summierung rechts über alle Zeilen erstreckt.

Setzt man, nach Analogie mit den Gleichungen (11), diese Größe

$$\mu_X^{(\mathfrak{z})^2} = \mu_X^2 \left(1 - \rho_{XY}^2\right), \tag{16}$$

so soll ρ_{XY} den Namen „Korrelationsverhältnis" von X in Bezug auf Y erhalten; es ergibt sich dafür aus eben diesem Ansatz die Bestimmung

$$\rho_{XY}^2 = 1 - \frac{\mu_X^{(\mathfrak{z})^2}}{\mu_X^2} \tag{17}$$

Der gleiche Vorgang, auf die Kolonnen angewendet, führt zu dem Korrelationsverhältnis von Y in Bezug auf X:

$$\rho_{YX}^2 = 1 - \frac{\mu_Y^{(\mathfrak{k})^2}}{\mu_Y^2} \tag{17*}$$

wobei durch das ($\mathfrak{k}$) im Zähler darauf hingewiesen wird, daß es sich um das gewogene Mittel der mittleren Abweichungsquadrate der Kolonnen handelt.

Wird dieser Vorgang auf eine Korrelationstafel angewendet mit sehr nahe geraden Hauptlinien, so fallen ρ_{XY}, ρ_{YX} nur wenig verschieden aus und stimmen mit dem Korrelationskoeffizienten r überein. Weichen sie untereinander (und von dem mitbestimmten r) erheblich ab, so ist das ein Hinweis auf krumme Hauptlinien, die um so stärker von Geraden sich unterscheiden, je größer die Differenz zwischen ρ_{XY} und r, bzw. ρ_{YX} und r ist.

Die Berechnung von $\mu_X^{(\mathfrak{Y})^2}$, bzw. $\mu_Y^{(\mathfrak{X})^2}$ kann, statt sie nach der Formel (15) und der analogen vorzunehmen, auf die Zeilen- und die Kolonnenmittel gegründet werden.

Aus der Figur liest man ab

$$x = x' + d;$$

wird dies quadriert und längs der Zeile summiert, so entsteht wegen $\Sigma x' = 0$

$$\Sigma (x^2) = \Sigma (x'^2) + n\, d^2 = n \left(\mu_X^{(Y)^2} + d^2\right)$$

und durch Summierung über alle Zeilen und Division durch N

$$\mu_X^2 = \mu_X^{(\mathfrak{Y})^2} + \mu_d^{(X)^2};$$

daraus erhält man schließlich, indem man durch μ_X^2 dividiert und (17) beachtet,

$$\rho_{XY}^2 = \frac{\mu_d^{(X)^2}}{\mu_X^2} \tag{18}$$

Dies gibt folgende Regel: Man bestimmt die Abweichungen der Zeilenmittel von M_X, bildet ihr quadratisches Mittel unter Berücksichtigung der Häufigkeiten und dividiert dieses durch die mittlere Abweichung aller X; dadurch erhält man das Korrelationsverhältnis von X in Bezug auf Y.

Ebenso ist mit den Kolonnen vorzugehen.

Man liest ferner an der Figur ab, daß

$$x'' = x' + \delta;$$

wird mit diesem Ansatz ebenso verfahren wie vorhin, so kommt man zu

$$m_X^2 = \mu_X^{(\mathfrak{Y})^2} + \mu_\delta^{(X)^2};$$

ersetzt man m_X^2 und $\mu_X^{(\mathfrak{Y})^2}$ durch ihre Werte aus den Gleichungen (11) und (16), so erhält man weiter

$$\mu_X^2 (1 - r^2) = \mu_X^2 \left(1 - \rho_{XY}^2\right) + \mu_\delta^{(X)^2},$$

woraus

$$\mu_\delta^{(X)^2} = \mu_X^2 \left(\rho_{XY}^2 - r^2\right). \tag{19}$$

Aus dieser Beziehung folgt mit Rücksicht auf den positiven Charakter der linken Seite, daß im allgemeinen $\rho_{XY}^2 > r^2$, wie oben schon im voraus angedeutet worden ist. Der Fall $\rho_{XY}^2 = r^2$ würde auf $\mu_\delta^{(X)^2} = 0$ führen, was nur dann eintreten kann, wenn die Zeilenmittel auf der Geraden Z liegen. Je größer aber $\rho_{XY}^2 - r^2$, um so größer ist auch $\mu_\delta^{(X)^2}$, um so mehr weichen die Zeilenmittel von einer Geraden ab.

82. Als Beispiel eignet sich in besonderem Maße die Tab. 57, an der bereits ein von der Geraden stark abweichender Verlauf der Kolonnenmittel festgestellt worden ist: es ist daher ein erheblicher Unterschied zwischen ρ_{YX} und r zu erwarten.

Die folgende Tabelle enthält die Daten und zeigt den Rechnungsgang.

Tab. 57a. Berechnung des Korrelationsverhältnisses ρ_{YX}.

Unter 100 Geborenen befinden sich Knaben: X	Mittel der betreffenden Kolonne $M_Y^{(X)}$	Abweichung vom allgem. Mittel $M_Y = 109{,}72$	Quadrat dieser Abweichung	Kolonnenhäufigkeit n	Produkt der beiden letzten Zahlen
30,5—32,0	37,50	72,22	5 215,73	1	5 215,73
32,0—33,5	37,50	72,22	5 215,73	1	5 215,73
33,5—35,0	.	.		.	.
35,0—36,5	12,50	97,22	9 451,73	1	9 451,73
36,5—38,0	62,50	47,22	2 229,73	1	2 229,73
38,0—39,5	48,21	61,51	3 783,48	7	26 484,36
39,5—41,0	54,17	55,55	3 085,80	3	9 257,40
41,0—42,5	50,00	59,72	3 566,48	10	35 664,80
42,5—44,0	50,83	58,89	3 468,03	15	52 020,45
44,0—45,5	146,59	— 36,87	1 359,40	11	14 953,40
45,5—47,0	92,76	16,96	287,64	19	5 465,16
47,0—48,5	116,79	— 7,07	49,98	35	1 749,30
48,5—50,0	153,02	— 43,30	1 874,89	29	54 371,81
50,0—51,5	131,62	— 21,90	479,61	34	16 306,74
51,5—53,0	196,67	— 86,95	7 560,30	30	226 809,00
53,0—54,5	108,33	1,39	1,93	30	57,90
54,5—56,0	104,55	5,17	26,73	22	588,06
56,0—57,5	100,46	9,26	85,75	27	2 315,25
57,5—59,0	68,27	41,45	1 718,10	13	22 335,30
59,0—60,5	79,17	30,55	933,30	12	11 199.60
60,5—62,0	65,00	44,72	1 999,88	10	19 998,80
62,0—63,5	75,00	34,72	1 205,48	2	2 410,96
63,5—65,0	45,83	63,89	4 081,93	6	24 491,58
65,0—66,5	37,50	72,22	5 215,73	2	10 431,46
66,5—68,0	37,50	72,22	5 215,73	3	15 647,19
				324	574 671,44

Aus den Schlußzahlen erhält man

$$\mu_d^{(Y)^2} = \frac{574\,671{,}44}{324} = 1773{,}68, \quad \mu_d^{(Y)} = 42{,}12$$

$$\rho_{YX} = \frac{42{,}12}{109{,}53} = 0{,}38:$$

das weicht in der Tat von dem zu der Tabelle errechneten Korrelationskoeffizienten $r = -0{,}009$ sehr erheblich ab.

Auch darin bestätigt sich die Theorie, daß bei der in Rede stehenden Tafel ρ_{XY} kleiner ist als ρ_{YX}, entsprechend der weit besseren Anpassung der Zeilenmittel an eine Gerade; es findet sich $\rho_{XY} = 0{,}21$.

Das Pearsonsche Korrelationsverhältnis hat E. Schäfer[1]) in der Marktanalyse angewandt, und zwar bei der Untersuchung des Zusammenhanges zwischen der Steuerkraft und der Zahl der Elektroapparate. Die Steuerkraft eines Gebiets mißt Schäfer durch die Kopfquote der veranlagten Einkommensteuer zuzüglich Vermögenssteuer, und die relative Häufigkeit der Elektroapparate bestimmt er durch Inbeziehungsetzen der absoluten Zahl der Elektroapparate zur Zahl der elektrifizierten Haushaltungen. Die sich ergebende Regressionslinie, die stark von einer geraden Linie abweicht, stellt er durch eine Kurve von der Form $Y = a + b \cdot \log X$ dar. Er findet, daß das mittels dieser Regressionslinie berechnete Korrelationsverhältnis für die Praxis noch wesentlich geeigneter ist als der Korrelationskoeffizient.

§ 7. Gebrauch des Korrelationskoeffizienten.

83. Aus dem bisher Vorgeführten geht die Bedeutung des Korrelationskoeffizienten für vergleichende Untersuchungen auf den verschiedensten Gebieten hervor; an einer späteren Stelle (Art. 95) wird auch einiges über seine praktische Verwendung zu sagen sein. In diesem Paragraphen sollen einige Probleme besprochen werden, bei deren Lösung Korrelationskoeffizienten als Rechnungsgrößen auftreten.

Wir beginnen mit der folgenden Frage.

Es ist die mittlere Abweichung der Summe mehrerer Variablen $X_1, X_2, \ldots X_n$ von ihrem arithmetischen Mittel zu bestimmen, wenn jede der Variablen durch eine Kollektivreihe gegeben ist.

Hierbei sind zwei Fälle zu unterscheiden, je nachdem die Variablen verbunden oder unverbunden sind in dem Sinne, daß zu jedem Werte von X_1 ein bestimmter Wert von X_2, von X_3 usw. gehört oder daß die Einzelwerte der Variablen beliebig miteinander kombiniert werden dürfen.

Es genügt, daß man sich bei Ableitung auf zwei Variable beschränke.

a) Bei Verbundenheit haben X_1 und X_2 gleiche Umfänge und denselben Umfang N hat auch $X = X_1 + X_2$. Sind M_1, M_2 die arithmetischen Mittel von X_1, X_2, so können die Einzelwerte in der Form $M_1 + x_1$, $M_2 + x_2$ geschrieben werden, die Summe erscheint dann in der Form $M_1 + M_2 + x_1 + x_2$, und da $M_1 + M_2$ das arithmetische Mittel vom X ist, so ist $x_1 + x_2$ die Abweichung x dieses Einzelwertes von X von seinem arithmetischen Mittel; nun ist

$$x^2 = x_1^2 + x_2^2 + 2 x_1 x_2$$
$$\Sigma(x^2) = \Sigma(x_1^2) + \Sigma(x_2^2) + 2\,\Sigma(x_1 x_2),$$

daraus folgt durch Division mit N:

$$\mu^2 = \mu_1^2 + \mu_2^2 + 2\frac{\Sigma(x_1 x_2)}{N};$$

[1]) E. Schäfer, Die Verbreitung von Elektro- und Gasapparaten. Eine marktanalytische Studie über die Absatzbedingungen in den 20 Verwaltungsbezirken Groß-Berlins. Stuttgart 1933, S. 49.

nun ist

$$r_{12} = \frac{\Sigma (x_1 x_2)}{N \mu_1 \mu_2}$$

der Korrelationskoeffizient zwischen X_1 und X_2, mithin ist μ^2 in der Form zu schreiben:

$$\mu^2 = \mu_1^2 + \mu_2^2 + 2 r_{12} \mu_1 \mu_2. \quad (1)$$

Steht statt der Summe die Differenz $X_1 - X_2$ in Frage, so verläuft die Ableitung genau in derselben Weise, nur mit dem Unterschiede, daß das letzte Glied negativ ausfällt. Somit gilt für die mittlere Abweichung der Differenz die Formel

$$\mu^2 = \mu_1^2 + \mu_2^2 - 2 r_{12} \mu_1 \mu_2. \quad (2)$$

Sind die Variablen korrelationslos, so ist $r_{12} = 0$, und beide Formeln gehen in die eine über:

$$\mu^2 = \mu_1^2 + \mu_2^2. \quad (3)$$

In dem anderen Grenzfalle, bei vollständiger Korrelation, u. zw. bei $r_{12} = +1$, wird

bei der Summe .... $\mu = \mu_1 + \mu_2$
bei der Differenz .... $\mu = |\mu_1 - \mu_2|$,

bei $r_{12} = -1$ hingegen

bei der Summe .... $\mu = |\mu_1 - \mu_2|$
bei der Differenz .... $\mu = \mu_1 + \mu_2$;

in allen zwischenliegenden Fällen ist $|\mu_1 - \mu_2| < \mu < \mu_1 + \mu_2$.

Die Ausdehnung auf beliebig viele Variable unterliegt keiner Schwierigkeit. man findet bei n Variablen

$$\mu^2 = \mu_1^2 + \mu_2^2 + \ldots + \mu_n^2 + 2 r_{12} \mu_1 \mu_2 + 2 r_{13} \mu_1 \mu_3 + \ldots + 2 r_{23} \mu_2 \mu_3 + \ldots ; \quad (4)$$

dabei bleibt der erste Teil der rechten Seite, die Quadratsumme, derselbe, mit welchem Zeichen auch die Variablen verbunden sind; von den doppelten Produkten aber werden diejenigen negativ, deren zugehörige Variable ungleich bezeichnet sind.

b) Sind die Variablen X_1, X_2 unverbunden, N_1, N_2 die Umfänge ihrer Wertreihen, so können die $N_1 N_2 = N$ Werte von x wie folgt angeordnet werden:

$$\begin{array}{lll} x_1^{(1)} + x_2^{(1)} & x_1^{(1)} + x_2^{(2)} \ldots & x_1^{(1)} + x_2^{(N_2)} \\ x_1^{(2)} + x_2^{(1)} & x_1^{(2)} + x_2^{(2)} \ldots & x_1^{(2)} + x_2^{(N_2)} \\ \ldots\ldots & & \\ x_1^{(N_1)} + x_2^{(1)} & x_1^{(N_1)} + x_2^{(2)} \ldots & x_1^{(N_1)} + x_2^{(N_2)}; \end{array}$$

somit wird

$$\Sigma (x^2) = N_2 \Sigma (x_1^2) + N_1 \Sigma (x_2^2) + 2 \Sigma (x_1 x_2).$$

Wegen der Unverbundenheit ist $\Sigma(x_1 x_2) = \Sigma(x_1)\Sigma(x_2) = 0$, weil x_1, x_2 die Abweichungen von den bezüglichen arithmetischen Mitteln sind. Die Division mit N ergibt also

$$\frac{\Sigma(x^2)}{N} = \frac{\Sigma(x_1^2)}{N_1} + \frac{\Sigma(x_2^2)}{N}, \text{ d. i. } \mu^2 = \mu_1^2 + \mu_2^2, \tag{5}$$

und dies gilt somit auch für die Differenz.

Für beliebig viele unverbundene Variable ergibt sich auf dieselbe Weise

$$\mu^2 = \mu_1^2 + \mu_2^2 + \ldots + \mu_n^2, \tag{6}$$

mit welchen Vorzeichen sie zur Summe verknüpft sein mögen.

Der Fall a) hätte z. B. zur Anwendung zu kommen, wenn an einer Reihe von Exemplaren eines Lebewesens die Gesamtausdehnung durch Messung und nachträgliche Summierung von Teildimensionen abgeleitet werden soll, die untereinander in Korrelation stehen; der Fall b) tritt ein, wenn für unabhängige Größen Beobachtungsreihen vorliegen und nach einer algebraischen Summe der Größen gefragt wird.

84. Mittlere Abweichung des arithmetischen Mittels. Es seien X_1, X_2, ... X_n Beobachtungen einer und derselben Größe X, M ihr arithmetisches Mittel, x_1, x_2, ... x_n, x die wahren Fehler von X_1, X_2, ... X_n, M, so besteht die Beziehung

$$M + x = \frac{X_1 + x_1 + X_2 + x_2 + \ldots + X_n + x_n}{n},$$

aus der

$$n x = x_1 + x_2 + \ldots + x_n$$

folgt; nun sei N_i die Zahl der verschiedenen Werte, deren x_i fähig ist, dann ist bei Unabhängigkeit der Beobachtungen $N = N_1 N_2 \ldots N_n$ die Zahl der Werte, die x annehmen kann, entsprechend den ebensovielen Verbindungen der Einzelwerte von x_1, x_2, .. x_n; mithin ergibt sich

$$n^2 \Sigma(x^2) = \frac{N}{N_1}\Sigma(x_1^2) + \frac{N}{N_2}\Sigma(x_2^2) + \ldots$$
$$+ \frac{N}{N_n}\Sigma(x_n^2) + 2\frac{N}{N_1 N_2}\Sigma(x_1 x_2) + 2\frac{N}{N_1 N_3}\Sigma(x_1 x_3) + \ldots;$$

denn das Quadrat eines Einzelwerts von x_i kommt so oft vor, als es Wertverbindungen der Einzelwerte der übrigen gibt und ebenso jedes Produkt zweier Einzelwerte von x_i, x_k. Wegen der Unabhängigkeit und daher auch Unverbundenheit ist

$$\Sigma(x_i x_k) = \Sigma(x_i)\Sigma(x_k);$$

sind nun alle Fehler von solcher Beschaffenheit, daß ihre positiven und negativen Einzelwerte sich das Gleichgewicht halten, so ist jedes $\Sigma(x_i) = 0$, somit auch jedes $\Sigma(x_i x_k) = 0$; es verbleibt also nach Division durch N

$${}^2\mu_M^2 = \mu_1^2 + \mu_2^2 + \ldots + \mu_n^2 \tag{7}$$

und damit ist μ_M bestimmt, wenn man μ_1, μ_2, ... μ_n kennt; es ist dann der mittlere Fehler des arithmetischen Mittels gleich der Quadratwurzel aus der Summe der mittleren Fehlerquadrate der einzelnen Beobachtungen, dividiert durch ihre Anzahl.

Von besonderem Interesse ist der Fall, daß die Beobachtungen gleichgenau sind, was mathematisch seinen Ausdruck darin finden soll, daß ihre mittleren Fehler übereinstimmen: $\mu_1 = \mu_2 = \ldots = \mu_n = \mu$; der letzte Ansatz geht dann über in

$$n \mu_M^2 = \mu^2,$$

woraus sich

$$\mu_M = \frac{\mu}{\sqrt{n}} \tag{8}$$

ergibt.

Diese Formel von fundamentaler Bedeutung besagt: Der mittlere Fehler des arithmetischen Mittels von n gleichgenauen Beobachtungen ergibt sich durch Division des mittleren Fehlers einer Beobachtung durch die Quadratwurzel aus n.

Dieser Begriff wird auch auf Kollektive übertragen; zu den beiden Größen, die bisher zur generellen Kennzeichnung eines Kollektivs vorzugsweise verwendet worden sind, nämlich zu dem arithmetischen Mittel M und der mittleren Abweichung μ der Kollektivglieder tritt als dritte Größe die mittlere Abweichung μ_M des arithmetischen Mittels. μ_M wird bestimmt nach der Formel (8), wenn n der Umfang des Kollektivs ist. Sowie μ ein Maß für die Streuung oder Variabilität der Kollektivglieder, so ist μ_M ein Maß für die Variabilität des aus ihnen gewonnenen Mittels. Je umfangreicher das Kollektiv, um so stabiler sein Mittelwert.

85. Um Belege zur Anwendung der vorstehenden Formeln, insbesondere der Formeln (3) und (8), zu geben, behandeln wir nachstehend zwei

Beispiele. 1) Aus den Verteilungstafeln der Körpergröße von 6194 Engländern und 1304 Schotten (Art. 66) ergibt sich

$$M_1 = 67{,}37 \qquad \mu_1 = 2{,}57 \text{ engl. Zoll}$$
$$M_2 = 68{,}61 \qquad \mu_2 = 2{,}50 \quad \text{,,} \quad \text{,,}$$

Welcher Schluß ist daraus auf den Größenunterschied der beiden Volksstämme zu ziehen?

Man wird den Größenunterschied naturgemäß nach den arithmetischen Mitteln beurteilen und daher zunächst sagen, daß die Schotten im Mittel um $d = 1{,}24$ Zoll größer seien als die Engländer. Aber der Unterschied zweier Mittel ist nicht bei jeder Größe gesichert, das hängt vielmehr von der Verläßlichkeit ab, mit welcher die Mittel selbst bestimmt sind, und diese ist nach ihren mittleren Abweichungen zu beurteilen. Aus ihnen ergibt sich die mittlere Abweichung der Differenz d

$$\sqrt{\frac{2{,}57^2}{6194} + \frac{2{,}50^2}{1304}} = 0{,}077 \text{ engl. Zoll},$$

und diese ist sehr klein im Vergleich zur Differenz d selbst. Eine Regel, die an einer späteren Stelle unter gewissen Voraussetzungen begründet werden wird (Art. 124), besagt, daß eine Differenz als „gesichert“ zu betrachten ist, wenn sie ihre mittlere Abweichung um das Drei- oder ein Mehrfaches übertrifft. Das trifft hier zu und darum kann man behaupten, daß die Schotten die Engländer tatsächlich in dem angegebenen Maße (bis auf dessen Unsicherheit) an Größe übertreffen.

2) Aus den mitgeteilten Verteilungstafeln der Gewichte neugeborener Knaben und Mädchen (Art. 25, 2) ergeben sich folgende Daten:

$$N_1 = 288; \qquad M_1 = 3278{,}1, \qquad \mu_1 = 474 \text{ g}$$
$$N_2 = 269; \qquad M_2 = 3092{,}6, \qquad \mu_2 = 447 \text{ „}$$

daraus berechnet man

$$\mu_{M_1} = 29{,}4 \qquad \mu_{M_2} = 27{,}3 \text{ g}$$

und den mittleren Fehler der Differenz $d = M_1 - M_2 = 185{,}5$

$$\mu_d = \sqrt{864{,}36 + 745{,}29} = 40{,}1 \text{ g}.$$

Da d 4,6 mal so groß ist wie μ_d, so kann der begründete Schluß gezogen werden, daß die Mädchen bei der Geburt durchschnittlich ein geringeres Gewicht haben als die Knaben, und zwar beträgt nach dem vorliegenden Material ihr durchschnittliches Gewicht 94,3% des durchschnittlichen Knabengewichtes (nach der Galtonschen Regel Art. 67 wären es 92,3%).

86. Wir stellen jetzt das folgende allgemeine Problem zur Lösung: X_1, X_2, X_3, ... seien miteinander verbundene Variable und $V = f(X_1, X_2, X_3, \ldots)$ eine beliebige Funktion derselben. Es seien N Wertverbindungen der Variablen beobachtet worden, so sind damit auch N Werte von V gegeben. Verlangt wird das arithmetische Mittel dieser Werte und ihre mittlere Abweichung von diesem, alles unter der Voraussetzung, daß die Abweichungen x_1, x_2, x_3, ... der Einzelwerte von X_1, X_2, X_3, ... von den zugehörigen Mittelwerten M_1, M_2, M_3, ... durchwegs klein sind im Vergleich zu diesen.

Den gesuchten Mittelwert bezeichnen wir mit M, die verlangte mittlere Abweichung mit μ.

Mit Rücksicht auf die getroffene Voraussetzung beschränken wir die Entwicklung von

$$V = f(M_1 + x_1,\ M_2 + x_2,\ M_3 + x_3, \ldots)$$

auf Glieder zweiter Ordnung und erhalten demgemäß, wenn $f(M_1, M_2, M_3, \ldots)$ mit f, seine Ableitungen in üblicher Weise mit $f_1, f_2, f_3, \ldots f_{11}, f_{22}, f_{33}, \ldots f_{12}, f_{13}, \ldots$ bezeichnet werden:

$$V = f + f_1 x_1 + f_2 x_2 + \cdots + \frac{1}{2}(f_{11} x_1^2 + f_{22} x_2^2 + \cdots) + f_{12} x_1 x_2 + f_{13} x_1 x_3 + \cdots;$$

daraus folgt

$$\left.\begin{aligned} M = \frac{1}{N}\Big[Nf + f_1 \Sigma(x_1) + f_2 \Sigma(x_2) + \cdots + \frac{1}{2}\big(f_{11}\Sigma(x_1^2) + f_{22}\Sigma(x_2^2) + \cdots\big) + \\ + f_{12}\Sigma(x_1 x_2) + f_{13}\Sigma(x_1 x_3) + \cdots\Big] = \\ = f + \frac{1}{2}(f_{11}\mu_1^2 + f_{22}\mu_2^2 + \cdots) + f_{12}\mu_1\mu_2 r_{12} + f_{13}\mu_1\mu_3 r_{13} + \cdots \end{aligned}\right\} \quad (9)$$

weil ja $\Sigma(x_1) = 0$, $\Sigma(x_2) = 0$ usw.

Um zu μ zu gelangen, ergibt sich analog der Beziehung $M^2 + \mu^2 = \frac{1}{N}\Sigma X^2$ (Art. 59) die Gleichung

$$M^2 + \mu^2 = \frac{1}{N}\Sigma(V^2). \qquad (10)$$

Bildet man V^2 unter derselben Einschränkung, daß man nämlich bei den Gliedern zweiten Grades stehen bleibt, so wird

$$\begin{aligned} V^2 = f^2 + 2f(f_1 x_1 + f_2 x_2 + \cdots) + f_1^2 x_1^2 + f_2^2 x_2^2 + \cdots + \\ + 2(f_1 f_2 x_1 x_2 + f_1 f_3 x_1 x_3 + \cdots) + f(f_{11} x_1^2 + f_{22} x_2^2 + \cdots) + \\ + 2f(f_{12} x_1 x_2 + f_{13} x_1 x_3 + \cdots), \end{aligned}$$

daraus weiter

$$\begin{aligned} \Sigma(V^2) = Nf^2 + f_1^2 \Sigma(x_1^2) + f_2^2 \Sigma(x_2^2) + \cdots + f\left(f_{11} \Sigma(x_1^2) + f_{22} \Sigma(x_2^2) + \cdots\right) + \\ + 2\left(f_1 f_2 \Sigma(x_1 x_2) + f_1 f_3 \Sigma(x_1 x_3) + \cdots\right) + \\ + 2f\left(f_{12} \Sigma(x_1 x_2) + f_{13} \Sigma(x_1 x_3) + \cdots\right) \end{aligned}$$

und

$$\begin{aligned} \frac{1}{N} \Sigma(V^2) = f^2 + f_1^2 \mu_1^2 + f_2^2 \mu_2^2 + \cdots + f(f_{11} \mu_1^2 + f_{22} \mu_2^2 + \cdots) + \\ + 2(f_1 f_2 \mu_1 \mu_2 r_{12} + f_1 f_3 \mu_1 \mu_3 r_{13} + \cdots) + \\ + 2f(f_{12} \mu_1 \mu_2 r_{12} + f_{13} \mu_1 \mu_3 r_{13} + \cdots); \end{aligned}$$

trägt man dies in (10) ein und rechnet aus dieser Gleichung mit Hilfe von (9) μ^2 aus, immer den gleichen Genauigkeitsgrad einhaltend, so erhält man

$$\mu^2 = f_1^2 \mu_1^2 + f_2^2 \mu_2^2 + \cdots + 2(f_1 f_2 \mu_1 \mu_2 r_{12} + f_1 f_3 \mu_1 \mu_3 r_{13} + \cdots). \quad (11)$$

Sind die Variablen unkorreliert, so geht die Lösung wegen $r_{12} = 0$, $r_{13} = 0$, ... über in

$$\left.\begin{aligned} M = f + \frac{1}{2}(f_{11} \mu_1^2 + f_{22} \mu_2^2 + \cdots) \\ \mu^2 = f_1^2 \mu_1^2 + f_2^2 \mu_2^2 + \cdots \end{aligned}\right\} \quad (12)$$

Die allgemeinen Formeln (9) und (11), auf die besonderen Fälle $V = X_1 X_2$ und $V = \frac{X_1}{X_2}$ angewendet, führen zu den folgenden Resultaten:

a) Bei $f = M_1 M_2$ ergibt sich $f_1 = M_2$, $f_2 = M_1$, $f_{11} = 0$, $f_{22} = 0$, $f_{12} = 1$ und hiermit wird

$$\begin{aligned} M = M_1 M_2 + \mu_1 \mu_2 r_{12} \\ \mu^2 = M_2^2 \mu_1^2 + M_1^2 \mu_2^2 + 2 M_1 M_2 \mu_1 \mu_2 r_{12}; \end{aligned}$$

führt man die Hilfsgrößen $\frac{\mu_1}{M_1} = \alpha_1$, $\frac{\mu_2}{M_2} = \alpha_2$ ein, die nach unserer Voraussetzung auch kleine Beträge vorstellen, so schreibt sich das Formelpaar, zur praktischen Ausführung besser geeignet:

$$\left.\begin{aligned} M = M_1 M_2 (1 + \alpha_1 \alpha_2 r_{12}) \\ \mu^2 = M_1^2 M_2^2 (\alpha_1^2 + \alpha_2^2 + 2\alpha_1 \alpha_2 r_{12}). \end{aligned}\right\} \quad (13)$$

Wenn kein korrelativer Zusammenhang zwischen den Faktoren besteht, hat man also

$$M = M_1 M_2$$
$$\mu^2 = M_1^2 M_2^2 (\alpha_1^2 + \alpha_2^2).$$

b) Aus $f = \frac{M_1}{M_2}$ folgen $f_1 = \frac{1}{M_2}$, $f_2 = -\frac{M_1}{M_2^2}$, $f_{11} = 0$, $f_{22} = \frac{2 M_1}{M_2^3}$, $f_{12} = -\frac{1}{M_2^2}$;

hiermit findet man

$$M = \frac{M_1}{M_2} + \frac{M_1}{M_2^3} \mu_2^2 - \frac{\mu_1 \mu_2}{M_2^2} r_{12}$$

$$\mu^2 = \frac{\mu_1^2}{M_2^2} + \frac{M_1^2}{M_2^4} \mu_2^2 - 2 \frac{M_1}{M_2^3} \mu_1 \mu_2 r_{12}$$

und bei Benützung derselben Hilfsgrößen

$$\left.\begin{aligned} M &= \frac{M_1}{M_2}(1 + \alpha_2^2 - \alpha_1 \alpha_2 r_{12}) \\ \mu^2 &= \frac{M_1^2}{M_2^2}(\alpha_1^2 + \alpha_2^2 - 2 \alpha_1 \alpha_2 r_{12}); \end{aligned}\right\} \quad (14)$$

sind Zähler und Nenner unkorreliert, so wird

$$M = \frac{M_1}{M_2}(1 + \alpha_2^2), \text{ also stets } M > \frac{M_1}{M_2}, \text{ und}$$

$$\mu^2 = \frac{M_1^2}{M_2^2}(\alpha_1^2 + \alpha_2^2).$$

Beispiele: 1) Wenn auf Grund der Tab. 55 (Art. 74) das Verhältnis der Kinderzahl der Mutter zu jener der Tochter ermittelt werden soll, so hat man dazu folgende Daten zur Verfügung (Art. 79, 4):

$$M_1 = 5{,}90, \quad \mu_1 = 2{,}83 \qquad r_{12} = 0{,}213.$$
$$M_2 = 4{,}34, \quad \mu_2 = 2{,}97$$

Daraus berechnen sich

$$\frac{M_1}{M_2} = 1{,}3594 \qquad \alpha_1 = 0{,}4797 \qquad \alpha_2 = 0{,}6843$$

und hiermit

$$M = 1{,}901, \qquad \mu = 1{,}016.$$

Das durchschnittliche Verhältnis zwischen mütterlicher und töchterlicher Kinderzahl bestimmt sich also mit rund 1,9 und einer mittleren Abweichung 1,016 [1]).

[1]) Zu beachten ist, daß nur solche Mütter in Betracht kommen, die unter ihren Kindern mindestens eine Tochter haben.

2) Auf Grund der Tab. 56 (Art. 80), aus der sich

$$\begin{array}{ll} M_1 = 3490{,}56 & \mu_1 = 512{,}10 \text{ g} \\ M_2 = 574{,}28 & \mu_2 = 117{,}44 \text{ „} \end{array} \qquad r_{12} = 0{,}6237$$

berechnen, erhält man als durchschnittliches Verhältnis zwischen Gewicht des (männlichen) Neugeborenen und der Plazenta 6,219 mit der mittleren Abweichung 0,978.

87. Es kommt vor, daß mehrere Teilkollektive ihrer Natur nach zu einem Gesamtkollektiv gehören. Wenn dann die arithmetischen Mittel für die einzelnen Teilkollektive bekannt sind, so entsteht die Frage, wie daraus das arithmetische Mittel für das Gesamtkollektiv abzuleiten ist.

Mitunter wird die Sache einfach so erledigt, daß man aus den Teilmitteln wieder das Mittel nimmt; dieser Vorgang führt aber im allgemeinem zu keinem brauchbaren Resultat, nur dann wäre er berechtigt, wenn man wüßte, daß den einzelnen Mittelwerten gleiches Gewicht zukommt. Da dies aber nur ganz ausnahmsweise der Fall sein wird, so müßten Daten zur Verfügung stehen, die eine Bestimmung oder doch eine Schätzung der Gewichte zulassen[1]).

Wenn z. B. der durchschnittliche Einheitspreis einer Ware auf verschiedenen Märkten eines Landes bekannt ist, so ist daraus der durchschnittliche Einheitspreis für das ganze Land nur dann ableitbar, wenn man auch die Mengen der betreffenden Ware, die auf den einzelnen Märkten verkauft worden sind, kennt. Diese Mengen sind nämlich die naturgemäßen „Gewichte", mit denen das gewogene Mittel aus den Marktmitteln zu bilden ist. Statt dessen wird in Ermanglung der ergänzenden Daten das gewöhnliche Mittel genommen, das sich aber von dem angestrebten Mittelwert mehr oder weniger unterscheiden wird.

Ein anderer Fall: Es sind die Geburtenziffern für die einzelnen Bezirke eines Landes bekannt. Die Geburtenziffer des ganzen Landes würde daraus hervorgehen durch Bildung eines gewogenen Mittels, wenn man die Bewohnerzahlen der Bezirke als Gewichte benützte. Das ungewogene Mittel gibt einen davon abweichenden Wert, dem keine Bedeutung zukommt.

Die Beziehung zwischen dem gewogenen und ungewogenen Mittel hängt von der Korrelation ab, die zwischen den Werten der Variablen und ihren Gewichten besteht.

Es heiße allgemein X die Variable, g ihr Gewicht; M das ungewogene, M_1 das gewogene arithmetische Mittel der beobachteten Einzelwerte von X, x ihre Abweichung von M, γ das mittlere Gewicht, μ_X die mittlere Abweichung der X von M, μ_g die mittlere Abweichung der g von γ. Dann bestehen die Gleichungen:

$$M = \frac{1}{N} \Sigma (X), \qquad M_1 = \frac{\Sigma (g X)}{\Sigma (g)}, \qquad \gamma = \frac{1}{N} \Sigma (g).$$

Nun ist

$$\begin{aligned} \Sigma (g X) &= \Sigma\, g\, (M + x) = M \Sigma (g) + \Sigma (g x) \\ &= N M \gamma + \Sigma (g x); \end{aligned}$$

[1]) Vgl. Art. 49.

ist r der Korrelationskoeffizient zwischen X und g, so hat man

$$\Sigma (g x) = N r \mu_X \mu_g,$$

mithin ist weiter

$$\Sigma (g X) = N M \gamma + N r \mu_X \mu_g,$$

und daraus folgt durch Division mit $\Sigma (g) = N \gamma$

$$M_1 = M + r \mu_X \frac{\mu_g}{\gamma}. \tag{11}$$

Man erkennt aus dieser Formel, daß das gewogene Mittel größer oder kleiner ist als das ungewogene, je nachdem die Korrelation zwischen X und g positiv oder negativ ist. Aber auch von der Größe des Verhältnisses $\frac{\mu_g}{\gamma}$ hängt der Unterschied ab: je beträchtlicher die Schwankungen in den Gewichten, um so größer ist er.

Unter günstigen Umständen kann die Formel (11) sogar zu einer indirekten Bestimmung des Korrelationskoeffizienten verwendet werden, indem

$$r = \frac{M_1 - M}{\mu_X} \cdot \frac{\gamma}{\mu_g}.$$

Ein Beispiel soll die Benützung dieser Formel erläutern. Zur Beurteilung der Verbreitung der Armut in England und Wales ist in den einzelnen Verwaltungsbezirken der Prozentsatz desjenigen Teils der Bevölkerung erhoben worden, der auf Grund des Armengesetzes Unterstützung empfängt. Die 632 Bezirke verteilten sich am 1. Jänner 1891 auf die verschiedenen Prozentsätze wie die nebenstehende Tabelle zeigt[1]).

Tab. 58.

Prozentsatz	Zahl der Bezirke
0,75—1,25	18
1,25—1,75	48
1,75—2,25	72
2,25—2,75	89
2,75—3,25	100
3,25—3,75	90
3,75—4,25	75
4,25—4,75	60
4,75—5,25	40
5,25—5,75	21
5,75—6,25	11
6,25—6,75	5
6,75—7,25	1
7,25—7,75	1
7,75—8,25	0
8,25—8,75	1
	632

Hieraus ergibt sich das ungewogene Mittel $M = 3{,}29$, ferner $\mu_X = 1{,}24\,\%_0$. Der aus der Gesamtbevölkerung abgeleitete Prozentsatz entspricht dem gewogenen Mittel und ergab sich mit $M_1 = 2{,}69$. Es fehlen zur Ausführung obiger Formel noch die Daten γ und μ_g. Nun fand sich für das Jahr 1891 die durchschnittliche Bevölkerung eines Bezirkes mit rund 45900 und die mittlere Abweichung in der Reihe der Bevölkerungszahlen, die in dem großen Spielraum zwischen 2000 und einer halben Million schwankten, mit 56400. Man bekommt also (es handelt sich um eine bloße Schätzung)

$$r = \frac{2{,}69 - 3{,}29}{1{,}24} \cdot \frac{45\,900}{56\,400} = -0{,}39,$$

eine negative Korrelation zwischen Prozentsatz und Bevölkerung.

[1]) G. U. Yule, An Introduction to the Theory of Statistics. London 1932, S. 139 u. 222.

Der Korrelationskoeffizient kann auch zur Kennzeichnung des Grades der Gleichförmigkeit einer Entwicklung benutzt werden. Zur Erfassung der Strukturwandlungen eines Kollektivs von einer Bestandserhebung bis zur nächsten hat S. Schott[1]) eine Methode angewandt, die er als „Statistik der beharrenden Fälle" bezeichnet. Schott gliedert die Gesamtheit der Haushaltungen, die in Mannheim bei den Zählungen 1916 und 1917 erfaßt wurden, zunächst in Gruppen nach der Zahl der Personen, die 1916 in den Haushaltungen lebten. Jede dieser Gruppen zerlegt er weiter in Untergruppen nach der Zahl der Personen, die 1917 zu den Haushaltungen gehörten. Diese Methode hat A. Pfütze[2]) unter der Bezeichnung „Entwicklungsstatistik" weiter ausgebaut. Pfütze legt seinen Untersuchungen die Bestandserhebungen über die gewerblichen Betriebe in Sachsen nach der Gewerbeaufsichtsstatistik in den Jahren 1926, 1928 und 1932 zugrunde und gliedert das Kollektiv der gewerblichen Betriebe, die 1926 und 1928 mehr als 30 Arbeitnehmer beschäftigen, zunächst nach der Zahl der beschäftigten Personen im Jahre 1926. Er bildet hierbei Gruppen mit der Gruppenbreite von 10 (Personen). Jede dieser Gruppen wird weiter nach der Personenzahl im Jahre 1928 in die gleichen Untergruppen aufgeteilt. Dasselbe nimmt er für die beiden Bestandserhebungen von 1928 und 1932 vor. Die Betriebsentwicklung von einer Erhebung zur anderen ist als vollkommen gleichförmig zu bezeichnen, wenn die Betriebe entweder alle in der gleichen Gruppe bleiben oder wenn sie um die gleiche Zahl von Gruppen auf- oder absteigen. In diesem Falle ist im statistischen Feld nur die Diagonale von links oben nach rechts unten oder eine Richtung, die zu dieser parallel läuft, besetzt. Erfolgt die Entwicklung bei den einzelnen Betrieben in verschiedener Stärke und auch in verschiedener Richtung, dann ist der durch das statistische Feld sich hindurchziehende Besetzungsstreifen breiter. Je breiter er ist, um so geringer ist der Grad der Gleichförmigkeit der Entwicklung. Die Breite des Besetzungsstreifens läßt sich zahlenmäßig durch den Korrelationskoeffizienten kennzeichnen. Daraus folgt, daß der Korrelationskoeffizient ein Maß für die Bestimmung des Grades der Gleichförmigkeit der Entwicklung ist. Im Falle der vollkommenen Gleichförmigkeit ist der Korrelationskoeffizient gleich + 1[3]).

§ 8. Korrelation zwischen mehr als zwei Variablen.

88. Die bisherigen Betrachtungen über Korrelation betrafen ausschließlich den Fall, daß es sich um die Abhängigkeit zweier Variablen handelt, wie dies etwa bei der Untersuchung der Fruchtbarkeit von Mutter und Tochter, Vater und Sohn, bei der Frage nach der Abhängigkeit von Länge und Breite der Blätter einer Pflanze u. ä. in die Erscheinung trat.

[1]) S. Schott, Ein Beitrag zur Statistik der beharrenden Fälle. Beiträge zur Statistik der Stadt Mannheim. Nr. 35, 1918.

[2]) A. Pfütze, Die Entwicklung gewerblicher Betriebe nach der Gewerbeaufsichtsstatistik von 1925 bis 1932. Zeitschrift des Sächsischen Statistischen Landesamtes, 78. und 79. Jahrgang, 1932 und 1933, S. 236 u. f.

[3]) Vgl. Betrachtungen über die entwicklungsstatistische Methode. Deutsches Statistisches Zentralblatt 1935, Heft 7, Sp. 197 u. f.

Nun ergeben sich auf verschiedenen Erscheinungsgebieten Fälle, wo mehr als zwei, also mindestens drei Variable in Beziehungen korrelativer Natur zueinander stehen. Man kann dann wohl die Korrelationen nach Paaren untersuchen, bekommt aber dadurch kein zutreffendes Bild des Zusammenhangs aller. Mit andern Worten, um das Zusammenwirken mehrerer Faktoren kennen zu lernen, genügt es nicht, sie paarweise zu kombinieren und ihre Abhängigkeit zu erforschen. So kann beispielsweise die bei einer Kulturpflanze erzielte Ernte von mehreren Faktoren abhängen, etwa von der Regenmenge und von der Temperatur während der maßgebenden Periode; dann treten drei Momente in Konkurrenz: Ernteertrag, Regenmenge und irgend ein Maß für die Wärmeverhältnisse. Es entstehen dann Fragen wie die folgenden: In welchem Maße hängt der Ernteertrag von der Regenmenge, von der Temperatur, von welchem Faktor mehr, von welchem weniger ab, wie beeinflussen sich Temperatur und Regenmenge gegenseitig? Um ein anderes Beispiel anzuführen: Man hat korrelative Beziehungen zwischen Enkeln und Großeltern festgestellt; wie hängen sie mit gleichartigen Beziehungen zwischen Eltern und Kindern zusammen?

Um Fragen solcher Art zu erledigen, ist es notwendig, alle in Betracht kommenden Variablen auf einmal in Beziehung zu setzen, da ja auch die ihnen entsprechenden Ursachen zusammen zur Wirkung kommen und nicht voneinander getrennt werden können.

Darin liegt ein Aufstieg der statistischen Forschung: erst richtet sich die Aufmerksamkeit auf eine einzelne Größe — damit wird der Grund gelegt für alle weitern Untersuchungen; dann werden zwei Größen, von welchen man weiß oder vermutet, daß sie voneinander abhängen, auf ihre Korrelation geprüft; schließlich wird ein ganzer Komplex einander beeinflussender Größen der Untersuchung unterzogen.

Bei diesem Fortschreiten gibt immer der vorherige Schritt die Richtung für den folgenden ab. So führt die Untersuchung der Korrelation zwischen zwei Variablen auf die Untersuchung der Verteilungen der einzelnen Variablen. Ebenso wird man für die Erforschung der Zusammenhänge zwischen mehreren Variablen eine Richtschnur suchen bei dem einfacheren Fall zweier Variablen.

89. Bei zwei Variablen bestand der wesentliche Gedanke darin, daß man dem Zusammenhang der Variablen eine lineare Gleichung zugrunde legte und deren noch unbestimmte Koeffizienten aus der Forderung ableitete, die Quadratsumme der Abweichungen solle ein Minimum werden. Daraus entsprangen die Korrelations- oder Regressionsgleichungen, denen geometrisch die Haupt- oder Regressionsgeraden entsprachen. In dem Korrelationskoeffizienten wurde ein Maß zur Beurteilung der Qualität und Stärke der Korrelation gefunden. Zugleich war damit der Hauptzweck erreicht, aus gegebener Änderung der einen Variablen auf die Änderung der andern schätzungsweise (im Durchschnitt) zu schließen. Es handelt sich jetzt darum, diese Gedankengänge auf mehrere Variable auszudehnen.

Es seien also $X_1, X_2, \ldots X_n$ die n Variablen, die für ein Erscheinungsgebiet wesentlich sind, zugleich in Betracht zu ziehen. Man nehme an, daß eine von ihnen, etwa X_1, als lineare Funktion der übrigen angenähert sich darstellen lasse, die so geschrieben werden möge:

$$X_1 = a + b_2 X_2 + b_3 X_3 + \ldots + b_n X_n. \tag{1}$$

Die Koeffizienten a, b_2, b_3, ... b_n als bereits bestimmt vorausgesetzt, wird die Annäherung darnach zu beurteilen sein, wie die Einsetzung zusammengehöriger Werte von X_1, X_2, ... X_n auf die Gleichung wirkt, d. h. um wieviel sich der Ausdruck

$$X_1 - a - b_2 X_2 - b_3 X_3 - \ldots - b_n X_n$$

von der Null entfernt, der er ja gleich sein müßte, wenn die lineare Beziehung streng gälte. Den Wert dieses Ausdruckes wollen wir als den „Fehler" der Gleichung für die betreffende Wertverbindung auffassen und ihm schon vorweg die Bezeichnung $x_{1.23\ldots n}$ geben, durch die erstens angezeigt wird, daß er sich auf die Bestimmung von X_1 bezieht und zweitens, daß dabei die Größen X_2, X_3, ... X_n zusammenwirken.

Jetzt aber stellen wir uns auf den Standpunkt, wie er tatsächlich vorliegt, daß nämlich die Koeffizienten unbestimmt sind und daß es sich darum handelt, für sie nach einem wissenschaftlich begründeten Prinzip geeignete Werte zu bestimmen. Als solches Prinzip benützen wir dasjenige, von dem sich die Methode der kleinsten Quadrate leiten läßt, und verlangen somit, daß die Summe der Fehlerquadrate

$$\Sigma\left(x^2_{1.23\ldots n}\right),$$

gebildet für alle Wertverbindungen, deren Anzahl N durch den Umfang des Kollektivs bezeichnet ist, ein Minimum erlangt.

Bevor wir an die Ausführung dieses Gedankens schreiten, soll auf die Rolle der Koeffizienten b hingewiesen werden. Aus dem Vorzeichen und dem Betrage von b_2 ist zu ersehen, in welchem Sinne und mit welcher Stärke X_2 auf X_1 einwirkt: ist b_2 positiv, so nimmt X_1 mit X_2 zu, im andern Falle ab; ist b_2 numerisch groß, so ist der Einfluß von X_2 auf X_1 stark, im andern Falle schwach. Wir nennen b_2 im Sinne der Galtonschen Terminologie den Regressionskoeffizienten von X_2 in Bezug auf X_1.

Um alle gegenseitigen Beziehungen zu erforschen, muß man jede der Variablen als Funktion der übrigen darstellen, also auch X_2 als Funktion von X_1, X_3, ... X_n, und zwar soll dies in derselben Form geschehen wie mit X_1, also durch eine Gleichung der Gestalt (1).

Dies erfordert die Einführung geeigneter symbolischer Bezeichnungen, aus welchen die Bedeutung der betreffenden Größe sofort ersichtlich ist.

90. Wir bezeichnen den Regressionskoeffizienten von X_2 in Bezug auf X_1 unter Mitberücksichtigung aller übrigen Variablen mit

$$b_{12.34\ldots n}$$

und nennen ihn einen partiellen Regressionskoeffizienten zum Unterschiede von dem totalen[1])

$$b_{12}.$$

[1]) Die Bezeichnung „partielle Korrelation" für die Fälle von mehr als zwei Variablen stammt von Pearson: sie ist dann auch auf die dabei auftretenden Korrelations- und Regressionskoeffizienten übertragen worden. Die mathematische Behandlung hat zuerst G. U. Yule gegeben (On the significance of Bravais formulae of regression etc., Proc. Roy. Soc., vol. 60, 1897).

der in Betracht käme, wenn X_1, X_2 allein ins Auge gefaßt würden. Die primären Zeiger 12 zeigen das Paar an, das in Beziehung gesetzt wird, die sekundären Zeiger weisen auf die mit ins Spiel kommenden Variablen hin. Der erste der beiden primären Zeiger kennzeichnet die darzustellende Variable. Es besteht also ein Unterschied zwischen $b_{12.34..n}$ und $b_{21.34..n}$; das erstemal wird X_1 durch X_2 und die andern, das zweitemal X_2 durch X_1 und die andern ausgedrückt. Die Ordnung der sekundären Zeiger ist ohne Belang, man wird sie in der natürlichen Ordnung schreiben.

Ist $x_1, x_2, \ldots x_n$ ein Wertsystem von $X_1, X_2, \ldots X_n$, das aus der Beobachtung hervorgegangen ist, so bezeichneten wir den Fehler, den es in der Gleichung (1) hervorruft, mit $x_{1.23\ldots n}$, wir haben also dafür den Ansatz:

$$x_{1.23\ldots n} = x_1 - a_1 - b_{12.3\ldots n}\, x_2 - b_{13.24\ldots n}\, x_3 - \ldots - b_{1n.23\ldots(n-1)}\, x_n. \quad (2)$$

Die Anzahl solcher Fehlergleichungen stimmt mit der Anzahl der beobachteten Wertverbindungen, also mit dem Umfang N des Kollektivs überein.

Der Fehlergleichungssysteme, wie ein solches durch (2) angedeutet ist, gibt es so viele, als Variable vorhanden sind, also n; das nächstfolgende hat die Gestalt:

$$x_{2.13\ldots n} = x_2 - a_2 - b_{21.3\ldots n}\, x_1 - b_{23.14\ldots n}\, x_3 - \ldots - b_{2n.13\ldots(n-1)}\, x_n, \quad (3)$$

u. s. w.

91. Die Bedingung, aus welcher die Regressionskoeffizienten des Systems (2) zu gewinnen sind, lautet

$$\Sigma(x^2_{1.23\ldots n}) = \Sigma(x_1 - a_1 - b_{12.3\ldots n}\, x_2 - b_{13.24\ldots n}\, x_3 - \ldots - b_{1n.23\ldots(n-1)}\, x_n)^2$$

ein Minimum. Führt man sie aus, indem man nach den Unbestimmten a_1 und den b einzeln differenziert und die Ableitungen gleich Null setzt, so ergeben sich die Gleichungen

$$\left.\begin{aligned}
&\Sigma(x_1 - a_1 - b_{12.3\ldots n}\, x_2 - b_{13.24\ldots n}\, x_3 - \ldots - b_{1n.23\ldots(n-1)}\, x_n) = 0 \\
&\Sigma x_2(x_1 - a_1 - b_{12.3\ldots n}\, x_2 - b_{13.24\ldots n}\, x_3 - \ldots - b_{1n.23\ldots(n-1)}\, x_n) = 0 \\
&\Sigma x_3(x_1 - a_1 - b_{12.3\ldots n}\, x_2 - b_{13.24\ldots n}\, x_3 - \ldots - b_{1n.23\ldots(n-1)}\, x_n) = 0 \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&\Sigma x_n(x_1 - a_1 - b_{12.3\ldots n}\, x_2 - b_{13.24\ldots n}\, x_3 - \ldots - b_{1n.23\ldots(n-1)}\, x_n) = 0,
\end{aligned}\right\} \quad (4)$$

die abgekürzt geschrieben werden können:

$$\left.\begin{aligned}
\Sigma(x_{1.23\ldots n}) &= 0 \\
\Sigma(x_2 \cdot x_{1.23\ldots n}) &= 0 \\
\Sigma(x_3 \cdot x_{1.23\ldots n}) &= 0 \\
&\ \vdots \\
\Sigma(x_n \cdot x_{1.23\ldots n}) &= 0
\end{aligned}\right\} \quad (4^*)$$

Ihre Zahl ist n, stimmt also mit der Zahl der Unbekannten überein. Nach der in der Methode der kleinsten Quadrate üblichen Benennung heißen diese Gleichungen Normalgleichungen.

Bislang ist über den Ausgangspunkt der Zählung der X keine Bestimmung getroffen worden; nun setzen wir fest, daß es die arithmetischen Mittel sein sollen, so daß $x_1, x_2, \dots x_n$ die Abweichungen der bezüglichen Werte der $X_1, X_2, \dots X_n$ von den Mittelwerten $M_1, M_2, \dots M_n$:

$$\begin{aligned} M_1 &= \frac{1}{N} \Sigma (X_1) \\ M_2 &= \frac{1}{N} \Sigma (X_2) \\ & \dots \\ M_n &= \frac{1}{N} \Sigma (X_n) \end{aligned} \tag{5}$$

sein sollen. Bei dieser Festsetzung ist

$$\begin{aligned} \Sigma (x_1) &= 0 \\ \Sigma (x_2) &= 0 \\ & \dots \\ \Sigma (x_n) &= 0. \end{aligned} \tag{6}$$

Hiermit ergibt die erste der Normalgleichungen das Resultat

$$a_1 = 0, \tag{7}$$

d. h. das konstante Glied der Regressionsgleichungen (1) und aller anderen ist gleich Null, sie haben also sämtlich die Form

$$x_1 = b_{12.3 \dots n} x_2 + b_{13.24 \dots n} x_3 + \cdots + b_{1n.23 \dots (n-1)} x_n. \tag{1*}$$

Die Auflösung der übrigen Normalgleichungen, aus denen die a ausgefallen sind, nach den b würde sich direkt sehr schwerfällig gestalten; es kämen dabei alle Produktsummen $\Sigma (x_1 x_2)$, $\Sigma (x_1 x_3)$, ... zur Anwendung. Es soll ein indirekter Weg eingeschlagen werden, der die Einführung verschiedener neuer Größen notwendig macht, trotzdem aber eine wesentliche Vereinfachung der Arbeit bedeutet.

92. In erster Linie soll der Begriff des Korrelationskoeffizienten auf den Fall von mehr als zwei Variablen ausgedehnt werden. Dies soll in einer Weise geschehen, die im Einklang steht mit dem Fall von bloß zwei Variablen X_1, X_2. Dort gab es nur zwei Regressionsgleichungen

$$x_1 = b_{12} x_2 \qquad x_2 = b_{21} x_1,$$

und es bestanden zwischen den Regressionskoeffizienten b_{12}, b_{21}, dem Korrelationskoeffizienten r_{12} und den mittleren Abweichungen μ_{X_1}, μ_{X_2} die Beziehungen (Art. 75, 6)

$$b_{12} = r_{12} \frac{\mu_{X_1}}{\mu_{X_2}} \qquad b_{21} = r_{12} \frac{\mu_{X_2}}{\mu_{X_1}},$$

aus welchen die von den μ freie Gleichung

$$r_{12}^2 = b_{12} \cdot b_{21}$$

resultiert. Dementsprechend soll der Korrelationskoeffizient zwischen X_1, X_2 unter Berücksichtigung der übrigen Variablen mit $r_{12.34\ldots n}$ bezeichnet und so festgesetzt werden, daß wieder

$$r_{12.34\ldots n}^2 = b_{12.34\ldots n} \cdot b_{21.34\ldots n} \tag{8}$$

besteht. Daß die rechte Seite der Vertauschung der Primärzeiger gegenüber invariant ist, stimmt zu der Tatsache, daß die Korrelation eine gegenseitige Beziehung ist, weshalb auch beim Korrelationskoeffizienten die Ordnung der Primärzeiger gleichgültig sein muß.

Des weiteren werde die Definition für die mittlere Abweichung einer Regressionsgleichung ebenso gestaltet wie bei zwei Variablen. Wenn es sich um die Form 1* handelt, sei das Zeichen $\mu_{1.23\ldots n}$ verwendet; das gibt

$$\mu_{1.23\ldots n}^2 = \frac{1}{N} \Sigma \left(x_{1.23\ldots n}^2\right). \tag{9}$$

Da die Summe rechts gebildet wird mit den aus den Normalgleichungen stammenden b-Werten, so stellt sie das Minimum dar.

Man spricht bei den Regressionskoeffizienten, den Korrelationskoeffizienten und den mittleren Abweichungen von einer Ordnung und meint darunter die Zahl der Sekundärzeiger; hiernach sind

$$b_{12.34\ldots n}, \qquad r_{12.34\ldots n}, \qquad \mu_{1.23\ldots n}$$

der Reihe nach von der Ordnung

$$n-2 \qquad n-2 \qquad n-1,$$

folgerichtig also die bisher gebrauchten Größen

$$b_{12} \qquad r_{12} \qquad \mu_1$$

von der Ordnung Null.

93. Wenn man den Bau der Normalgleichungen in der Form (4*) näher betrachtet, so kann man ihn als Ausdruck des folgenden Satzes hinstellen: Die Produktsummen aus den Abweichungen einer höheren Ordnung mit den Abweichungen nullter Ordnung, sofern der Zeiger der letzteren unter den Sekundärzeigern der ersteren vorkommt, sind durchwegs Null.

Dieser Satz soll allgemeine Geltung haben. Er bewährt sich in der Tat auch bei den Abweichungen zweiter Ordnung; denn aus

$$x_{1.2} = x_1 - b_{12} x_2 \qquad x_{2.1} = x_2 - b_{21} x_1$$

folgt

$$\Sigma (x_2 \cdot x_{1.2}) = \Sigma (x_1 x_2) - b_{12} \Sigma (x_2^2) \qquad \Sigma (x_1 \cdot x_{2.1}) = \Sigma (x_1 x_2) - b_{21} \Sigma (x_1^2),$$

und ersetzt man rechter Hand alles entsprechend den Formeln von Art. 75, so zeigt sich, daß $\Sigma (x_2 \cdot x_{1.2}) = 0$ und $\Sigma (x_1 \cdot x_{2.1}) = 0$ ist.

Die Anwendung des Satzes auf die Produktsumme

$$\Sigma (x_{1.34\ldots n} x_{2.34\ldots n})$$

gibt, wenn man einmal den ersten Faktor durch seinen Ausdruck

$$x_1 - b_{13.4\ldots n} x_3 - b_{14.35\ldots n} x_4 - \cdots - b_{1n.34\ldots(n-1)} x_n,$$

ein andermal den zweiten durch seinen Ausdruck

$$x_2 - b_{23.4\ldots n} x_3 - b_{24.35\ldots n} x_4 - \cdots - b_{2n.34\ldots(n-1)} x_n$$

ersetzt und die Summe entwickelt, die folgende Beziehung:

$$\Sigma (x_{1.34\ldots n} \cdot x_{2.34\ldots n}) = \Sigma (x_{1.34\ldots n} \cdot x_2) = \Sigma (x_1 \cdot x_{2.34\ldots n}),$$

die so ausgesprochen werden kann: Eine Produktsumme von Abweichungen bleibt ungeändert, wenn man die beiden Faktoren gemeinsamen Sekundärzeiger in dem einen oder andern Faktor fortläßt. Umgekehrt wird eine Produktsumme von der Form $\Sigma(x_1 \cdot x_{2.34\ldots n})$ nicht geändert, wenn man die im ersten Faktor fehlenden Sekundärzeiger durch jene des zweiten Faktors ergänzt.

Man kann hiernach die zweite Normalgleichung (4*) auch schreiben

$$\Sigma (x_{2.34\ldots n} \cdot x_{1.234\ldots n}) = 0,$$

und indem man den zweiten Faktor durch seinen Ausdruck (2) ersetzt, entwickelt und von den eben angeführten Eigenschaften der Produktsummen Gebrauch macht, entsteht die Gleichung

$$\Sigma (x_1 \cdot x_{2.34\ldots n}) - b_{12.34\ldots n} \Sigma (x_2 \cdot x_{2.34\ldots n}) = 0,$$

die auch geschrieben werden kann

$$\Sigma (x_{1.34\ldots n} \cdot x_{2.34\ldots n}) - b_{12.34\ldots n} \Sigma (x_{2.34\ldots n} \cdot x_{2.34\ldots n}) = 0,$$

woraus sich ergibt

$$b_{12.34\ldots n} = \frac{\Sigma (x_{1.34\ldots n} \cdot x_{2.34\ldots n})}{\Sigma (x^2_{2.34\ldots n})}. \tag{10}$$

Entsprechend den Gleichungen (3), (5) in Art. 75, die für den Fall zweier Variablen galten, hat man jetzt zu setzen

$$r_{12.34\ldots n} = \frac{\Sigma (x_{1.34\ldots n} \cdot x_{2.34\ldots n})}{N \mu_{1.34\ldots n} \mu_{2.34\ldots n}}$$

wobei die μ der Gleichung (9) entsprechend zu bilden sind; in Verbindung mit der vorangehenden Gleichung, in der der Nenner gleichbedeutend ist mit $N \mu^2_{2.34\ldots n}$, führt dies zu

$$b_{12.34\ldots n} = r_{12.34\ldots n} \frac{\mu_{1.34\ldots n}}{\mu_{2.34\ldots n}}. \tag{11}$$

Durch Vertauschung der Primärzeiger, die bei r keine Wirkung hat, wird daraus

$$b_{21.34\ldots n} = r_{12.34\ldots n} \frac{\mu_{2.34\ldots n}}{\mu_{1.34\ldots n}} \tag{11*}$$

und die Multiplikation beider Gleichungen führt wirklich auf die anfängliche Festsetzung (8) betreffend den Korrelationskoeffizienten.

94. Durch die Formel (11) wäre schon ein Weg zur Berechnung des Regressionskoeffizienten angegeben. Er läßt sich aber abkürzen durch den Umstand, daß sich Regressionskoeffizienten und mittlere Abweichungen einer bestimmten Ordnung zurückführen lassen auf solche der nächstniederen Ordnung. Auch das beruht auf den Eigenschaften der Produktsummen.

Es ist nämlich

$$\begin{aligned}\Sigma\left(x^2_{1.23\ldots n}\right) &= \Sigma\left(x_{1.23\ldots n} \cdot x_{1.23\ldots n}\right)\\ &= \Sigma\left(x_{1.23\ldots(n-1)}\left(x_1 - b_{1n.23\ldots(n-1)}\,x_n - \cdots\right)\right)\\ &= \Sigma\left(x_{1.23\ldots(n-1)} \cdot x_1\right) - b_{1n.23\ldots(n-1)}\,\Sigma\left(x_{1.23\ldots(n-1)} \cdot x_n\right),\end{aligned}$$

weil alle übrigen, aus $x_2, x_3, \ldots x_{n-1}$ hervorgehenden Summen gleich Null sind, und weiter

$$\begin{aligned}&= \Sigma\left(x_{1.23\ldots(n-1)} \cdot x_{1.23\ldots(n-1)}\right) -\\ &\quad - b_{1n.23\ldots(n-1)}\,\Sigma\left(x_{1.23\ldots(n-1)}\,x_{n.23\ldots(n-1)}\right)\\ &= \Sigma\left(x^2_{1.23\ldots(n-1)}\right) - b_{1n.23\ldots(n-1)}\,\Sigma\left(x_{1.23\ldots(n-1)} \cdot x_{n.23\ldots(n-1)}\right);\end{aligned}$$

die darin vorkommenden drei Summen können zufolge (9) und (10) der Reihe nach durch

$$N\mu^2_{1.23\ldots n} \qquad N\mu^2_{1.23\ldots(n-1)} \qquad N\mu^2_{1.23\ldots(n-1)}\,b_{n1.23\ldots(n-1)}$$

ersetzt werden; alsdann aber ergibt sich

$$\mu^2_{1.23\ldots n} = \mu^2_{1.23\ldots(n-1)}\left(1 - b_{1n.23\ldots(n-1)}\,b_{n1.23\ldots(n-1)}\right) \tag{12}$$

und mit Rücksicht auf (8)

$$\mu^2_{1.23\ldots n} = \mu^2_{1.23\ldots(n-1)}\left(1 - r^2_{1n.23\ldots(n-1)}\right). \tag{12*}$$

Aus dieser Formel ersieht man, daß auch alle Korrelationskoeffizienten höherer Ordnung echte Brüche sind. Zugleich ist damit eine Rekursionsformel gewonnen, die zu einer independenten Darstellung von $\mu^2_{1.23\ldots n}$ führt; im Sinne von (12*) bestehen nämlich weiter die Ansätze:

$$\mu^2_{1.23\ldots(n-1)} = \mu^2_{1.23\ldots(n-2)}\left(1 - r^2_{1(n-1).23\ldots(n-2)}\right)$$
$$\mu^2_{1.23\ldots(n-2)} = \mu^2_{1.23\ldots(n-3)}\left(1 - r^2_{1(n-2).23\ldots(n-3)}\right)$$

$$\mu^2_{1.2} \qquad = \mu^2_1\left(1 - r^2_{12}\right)$$

und durch Multiplikation aller mit Einschluß von (12*) kommt die Formel

$$\mu^2_{1.23\ldots n} = \mu^2_1\left(1 - r^2_{12}\right)\left(1 - r^2_{13.2}\right)\left(1 - r^2_{14.23}\right)\ldots\left(1 - r^2_{1n.23\ldots(n-1)}\right) \tag{13}$$

zustande.

In analoger Weise erhält man nach und nach

$$\Sigma\left(x_{1.34\ldots n}\cdot x_{2.34\ldots n}\right) =$$
$$= \Sigma\, x_{1.34\ldots(n-1)}\left(x_2 - b_{2n.34\ldots(n-1)}\, x_n - \ldots\right)$$
$$= \Sigma\left(x_{1.34\ldots(n-1)}\cdot x_2\right) - b_{2n.34\ldots(n-1)}\,\Sigma\left(x_{1.34\ldots(n-1)}\cdot x_n\right)$$
$$= \Sigma\left(x_{1.34\ldots(n-1)}\cdot x_{2.34\ldots(n-1)}\right) - b_{2n.34\ldots(n-1)}\,\Sigma\left(x_{1.34\ldots(n-1)}\cdot x_{n.34\ldots(n-1)}\right);$$

die drei Summen sind aber zufolge (10) ersetzbar durch

$$N\mu^2_{2.34\ldots n}\, b_{12.34\ldots n},\; N\mu^2_{2.34\ldots(n-1)}\, b_{12.34\ldots(n-1)},\; N\mu^2_{n.34\ldots(n-1)}\, b_{1n.34\ldots(n-1)};$$

dadurch kommt man zu der Gleichung

$$\mu^2_{2.34\ldots n}\, b_{12.34\ldots n}$$
$$= \mu^2_{2.34\ldots(n-1)}\, b_{12.34\ldots(n-1)} - \mu^2_{n.34\ldots(n-1)}\, b_{1n.34\ldots(n-1)}\, b_{2n.34\ldots(n-1)};$$

(11) und (11*) zufolge ist aber

$$\frac{b_{2n.34\ldots(n-1)}}{b_{n2.34\ldots(n-1)}} = \frac{\mu^2_{2.34\ldots(n-1)}}{\mu^2_{n.34\ldots(n-1)}}.$$

Man benützt diese Proportion, um aus der vorstehenden Gleichung $b_{2n.34\ldots(n-1)}$ zu eliminieren, und ersetzt die Größe $\mu_{2.34\ldots n}$ durch den Ausdruck

$$\mu^2_{2.34\ldots(n-1)}\left(1 - b_{2n.34\ldots(n-1)}\, b_{n2.34\ldots(n-1)}\right),$$

der analog der Gleichung (12) gebildet wird. Dann bleibt eine Gleichung zurück, die nur die b enthält und aus der sich die Rekursionsformel ergibt:

$$b_{12.34\ldots n} = \frac{b_{12.34\ldots(n-1)} - b_{1n.34\ldots(n-1)}\, b_{n2.34\ldots(n-1)}}{1 - b_{2n.34\ldots(n-1)}\, b_{n2.34\ldots(n-1)}}. \tag{14}$$

Ersetzt man im Zähler die b durch ihre Ausdrücke gemäß der Formel (11) und wendet auf den Nenner Formel (8) an, so wird

$$b_{12.34\ldots n} = \frac{r_{12.34\ldots(n-1)} - r_{1n.34\ldots(n-1)}\, r_{2n.34\ldots(n-1)}}{1 - r^2_{2n.34\ldots(n-1)}} \cdot \frac{\mu_{1.34\ldots(n-1)}}{\mu_{2.34\ldots(n-1)}};$$

daraus durch Vertauschung der Primärzeiger

$$b_{21.34\ldots n} = \frac{r_{12.34\ldots(n-1)} - r_{1n.34\ldots(n-1)}\, r_{2n.34\ldots(n-1)}}{1 - r^2_{1n.34\ldots(n-1)}} \cdot \frac{\mu_{2.34\ldots(n-1)}}{\mu_{1.34\ldots(n-1)}},$$

schließlich durch Multiplikation beider Ansätze und Ausziehung der Quadratwurzel

$$r_{12.34\ldots n} = \frac{r_{12.34\ldots(n-1)} - r_{1n.34\ldots(n-1)}\, r_{2n.34\ldots(n-1)}}{\left(1 - r^2_{1n.34\ldots(n-1)}\right)^{\frac{1}{2}} \left(1 - r^2_{2n.34\ldots(n-1)}\right)^{\frac{1}{2}}} \tag{15}$$

womit eine Rekursionsformel für die Korrelationskoeffizienten gefunden ist, deren wiederholte Anwendung schließlich bis auf die Koeffizienten nullter Ordnung hinführt, nämlich zu

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\left(1 - r^2_{13}\right)^{\frac{1}{2}} \left(1 - r^2_{23}\right)^{\frac{1}{2}}}. \tag{16}$$

95. Der Rechnungsgang bei Ausführung eines praktischen Falles wird der folgende sein.

Als Vorarbeit ist zu leisten die Bestimmung der arithmetischen Mittel M_1, M_2, M_3, ..., der mittleren Abweichungen μ_1, μ_2, μ_3, ... der Wertreihen von X_1, X_2, X_3, ... und der Korrelationskoeffizienten r_{12}, r_{13}, ... r_{23}, ... ihrer sämtlichen Paare. Damit sind die grundlegenden Daten gewonnen.

I. Man beginnt mit der Berechnung der Korrelationskoeffizienten höherer Ordnung, von der Formel (16) aufsteigend, gemäß (15).

II. Hierauf bestimmt man unter Verwendung der Formel (13) die mittleren Abweichungen der höchsten Ordnung.

III. Schließlich geht man an die Berechnung der notwendigen Regressionskoeffizienten, wozu sich die Formel (11) eignet.

Bei drei Variablen ist der ganze Formelapparat der folgende:

Ausgangsgrößen: M_1, M_2, M_3; μ_1, μ_2, μ_3; r_{12}, r_{13}, r_{23}.

I.

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\left(1-r^2_{13}\right)^{\frac{1}{2}} \left(1-r^2_{23}\right)^{\frac{1}{2}}},\quad r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{\left(1-r^2_{12}\right)^{\frac{1}{2}} \left(1-r^2_{23}\right)^{\frac{1}{2}}},\quad r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\left(1-r^2_{12}\right)^{\frac{1}{2}} \left(1-r^2_{13}\right)^{\frac{1}{2}}}.$$

II.

$$\mu_{1.23} = \mu_1 \left(1 - r_{12}^2\right)^{\frac{1}{2}} \left(1 - r_{13.2}^2\right)^{\frac{1}{2}}, \qquad \mu_{2.13} = \mu_2 \left(1 - r_{12}^2\right)^{\frac{1}{2}} \left(1 - r_{23.1}^2\right)^{\frac{1}{2}},$$

$$\mu_{3.12} = \mu_3 \left(1 - r_{13}^2\right)^{\frac{1}{2}} \left(1 - r_{23.1}^2\right)^{\frac{1}{2}}.$$

Wegen der Einflußlosigkeit der Ordnung der Sekundärzeiger kann jede dieser Größen noch auf eine zweite Art gerechnet werden, so z. B.

$$\mu_{1.23} = \mu_1 \left(1 - r_{13}^2\right)^{\frac{1}{2}} \left(1 - r_{12.3}^2\right)^{\frac{1}{2}},$$

worin eine Kontrolle der Rechnung liegt.

III.

$b_{12.3} = r_{12.3} \dfrac{\mu_{1.3}}{\mu_{2.3}}$. Diese Formel würde die Ausrechnung der mittleren Abweichungen erster Ordnung voraussetzen, die sonst keine weitere Verwendung finden. Es ist nach der zweiten Rechnungsart von $\mu_{1.23}$ in II

$$\mu_{1.23} = \mu_1 \left(1 - r_{13}^2\right)^{\frac{1}{2}} \left(1 - r_{12.3}^2\right)^{\frac{1}{2}}$$

$$\mu_{2.13} = \mu_2 \left(1 - r_{23}^2\right)^{\frac{1}{2}} \left(1 - r_{12.3}^2\right)^{\frac{1}{2}},$$

daraus

$$\frac{\mu_{1.23}}{\mu_{2.13}} = \frac{\mu_1}{\mu_2} \left(\frac{1 - r_{13}^2}{1 - r_{23}^2}\right)^{\frac{1}{2}};$$

anderseits hat man

$$\mu_{1.3} = \mu_1 \left(1 - r_{13}^2\right)^{\frac{1}{2}} \qquad \mu_{2.3} = \mu_2 \left(1 - r_{23}^2\right)^{\frac{1}{2}},$$

also ist auch

$$\frac{\mu_{1.3}}{\mu_{2.3}} = \frac{\mu_1}{\mu_2} \left(\frac{1 - r_{13}^2}{1 - r_{23}^2}\right)^{\frac{1}{2}},$$

daher

$$\frac{\mu_{1.3}}{\mu_{2.3}} = \frac{\mu_{1.23}}{\mu_{2.13}}.$$

Man hat also schließlich

$$b_{12.3} = r_{12.3} \frac{\mu_{1.23}}{\mu_{2.13}} \qquad b_{13.2} = r_{13.2} \frac{\mu_{1.23}}{\mu_{3.12}}$$

$$b_{21.3} = r_{12.3} \frac{\mu_{2.13}}{\mu_{1.23}} \qquad b_{23.1} = r_{23.1} \frac{\mu_{2.13}}{\mu_{3.12}}$$

$$b_{31.2} = r_{13.2} \frac{\mu_{3.12}}{\mu_{1.23}} \qquad b_{32.1} = r_{23.1} \frac{\mu_{3.12}}{\mu_{2.13}}$$

Hiermit sind alle drei Regressionsgleichungen bestimmt:

$$x_1 = b_{12.3}\, x_2 + b_{13.2}\, x_3$$
$$x_2 = b_{21.3}\, x_1 + b_{23.1}\, x_3$$
$$x_3 = b_{31.2}\, x_1 + b_{32.1}\, x_2.$$

Verlegt man den Nullpunkt der Zählung wieder nach dem Ursprung, was der Transformation

$$x_1 = X_1 - M_1 \qquad x_2 = X_2 - M_2 \qquad x_3 = X_3 - M_3$$

entspricht, so ergeben sich Gleichungen zwischen den ursprünglichen Variablen von der Form

$$X_1 = K_1 + b_{12.3}\, X_2 + b_{13.2}\, X_3$$
$$X_2 = K_2 + b_{21.3}\, X_1 + b_{23.1}\, X_3$$
$$X_3 = K_3 + b_{31.2}\, X_1 + b_{32.1}\, X_2.$$

In geometrischer Deutung sind es die Gleichungen dreier Ebenen, die durch den Punkt $M_1 | M_2 | M_3$ gehen. Diese Ebenen sind das räumliche Analogon jener zwei Geraden, die bei der zweidimensionalen Korrelation als deren Hauptlinien erkannt worden sind.

Für die Beurteilung der Verläßlichkeit der drei Gleichungen sind die Größen $\mu_{1.23}$, $\mu_{2.13}$, $\mu_{3.12}$ maßgebend. Von diesem Gesichtspunkte wäre diejenige zu bevorzugen, zu der die kleinste mittlere Abweichung gehört. Der praktische Gesichtspunkt kann aber von vornherein einer von ihnen den Vorzug geben, man wird sich dann auf die Entwicklung dieser einen beschränken, wodurch an Rechenarbeit gespart wird.

Für die logarithmische Rechnung, zu der man wohl stets greifen wird, sind die Formeln der Gruppen II, III unmittelbar geeignet, vorausgesetzt, daß man sich die darin vorkommenden Quadratwurzeln im voraus berechnet hat. Nur in der Gruppe I erfordert der Zähler eine besondere Arbeit.

96. Bevor an die Vorführung von Beispielen geschritten wird, sollen einige allgemeine Bemerkungen über die bisherige Anwendung der Korrelationstheorie eingeschaltet werden.

Den ausgedehntesten Gebrauch hat sie erfahren in der Biologie und in der Erblichkeitslehre; Vertreter dieser Richtungen waren ihre Begründer (Galton, Pearson). Neuerdings hat F. Lenz[1] die Korrelationsrechnung auf wichtige Probleme der menschlichen Erbforschung angewandt. Wertvolle Dienste leistet die Korrelationsrechnung bei der Bestimmung des Anteils von Erbanlage und Umwelt an der Entstehung bestimmter Eigenschaften[2]. Sehr wichtige Unterlagen liefert

[1]) F. Lenz, Menschliche Erblehre. 4. Auflage, Bd. I, München 1936, S. 627 u. f.

[2]) F. Lenz und O. v. Verschuer, Zur Bestimmung des Anteils von Erbanlage und Umwelt an der Variabilität. Archiv für Rassen- und Gesellschaftsbiologie 1928, Bd. 20, H. 4, S. 428.

hierzu die Zwillingsforschung. Lenz hat als erster gezeigt, wie die Berechnung von Korrelationen bei Zwillingen durchzuführen ist [1]).

Weiter sind mittels der Korrelationstheorie Probleme sozialer und wirtschaftlicher Natur, bei denen der praktische Gesichtspunkt der vorherrschende ist, behandelt worden [2]). Einige hierhergehörige Stoffe sollen kurz angedeutet werden, um Anregung zu solchen und ähnlichen Untersuchungen zu geben, wo das notwendige statistische Material vorhanden ist oder beschafft werden kann. Untersucht wurden die Zusammenhänge zwischen Heiratsziffer und Geburtenziffer; zwischen Heiratsziffer und Handelsverkehr; zwischen Heiratsziffer und Arbeitslosigkeit; zwischen Heiratsziffer und Weizenpreis; zwischen ehelicher Fruchtbarkeit (gemessen an der Zahl der Geburten, die auf 1000 verheiratete Frauen im gebärfähigen Alter kommen) und sozialer Lage (diese nach verschiedenen Momenten beurteilt, so nach dem relativen Anteil der Zahl der Dienenden, der Erwerbstätigen, der gewöhnlichen Arbeiter, der Leihhäuser im Vergleich zur Bevölkerung). Untersucht wurde ferner die Verbreitung und die Abnahme der Tuberkulose in ihrer Abhängigkeit von den verschiedenen Ursachen; der Einfluß der fortschreitenden Absonderung der Kranken in geschlossenen Heilstätten. Einen weiteren Gegenstand bildet die Untersuchung der Armutsverhältnisse in ihrer Abhängigkeit von verschiedenen Umständen, so von dem Verhältnis der häuslich Unterstützten (offene Fürsorge) zu den in Anstalten Untergebrachten (geschlossene Fürsorge), von dem relativen Anteil der über 65 Jahre alten Personen, von den Änderungen der Bevölkerungszahl selbst. Aus dem Bereich der Finanzstatistik wurden die Bewegungen in den Reserven der Banken, ihren Darlehen, den Einzahlungen, dem Diskont auf ihre Zusammenhänge geprüft.

Ausgedehnte Anwendung fand die Korrelationstheorie auf die Fragen der Produktionsstatistik. Die Zusammenhänge zwischen Ernteertrag und den meteorologischen Faktoren und die Beziehungen der letzteren untereinander sind von außerordentlicher Bedeutung; die hauptsächlichen Nutzpflanzen zeigen hierin verschiedenes Verhalten und waren Gegenstand vielfacher Untersuchung.

Bei vielen der angeführten Gegenstände zeigen sich zwei Arten der Änderung: eine langsam vor sich gehende säkulare (Trendbewegung) und neben ihr ein ständiges Schwanken, bestehend in kurzen oft beträchtlichen Ausschlägen bald nach der einen, bald nach der andern Seite (Saison- und Konjunkturschwankungen). Diese letztere Art der Änderung ist es zumeist, welche aktuelle Bedeutung hat. Um sich von der säkularen Bewegung unabhängig zu machen, hat man statt der individuellen Werte der betreffenden Variablen ihre „momentanen Mittelwerte" herangezogen, nämlich die Durchschnittswerte aus 3, 5, 7, 9 aufeinander folgenden Jahren, in deren Mitte das eben betrachtete steht. Die Abweichungen von diesen Mittelwerten erwiesen sich häufig als brauchbarer als die Werte selbst [3]).

Es hat sich weiter ergeben, daß nicht immer „gleichzeitige" Werte der in Beziehung gesetzten Variablen den Zusammenhang hervortreten lassen, daß viel-

[1]) F. Lenz, Handbuch der hygienischen Untersuchungsmethoden, herausgegeben von Gotschlich, Bd. 3, Jena 1929, S. 689 u. f.

[2]) Yule hat darüber der Tagung des Internationalen Statistischen Instituts in Paris 1909 einen Bericht (The Applications of the Method of Correlation to Social and Economic Statistics) vorgelegt, dem ein Verzeichnis der bezüglichen Literatur angeschlossen ist (Bulletin de l'Institut international de Statistique, Bd. 13, 1909, S. 537).

[3]) Vgl. Art. 99.

Tab. 59.

Korrelations-koeffizienten 0ter Ordnung $r_{\alpha\beta}$	$\log\left(1-r_{\alpha\beta}^2\right)^{\frac{1}{2}}$	Produktglied des Zählers	Zähler	log | Zähler |	log Nenner	Korrelationskoeffizienten 1. Ordnung: $\log \| r_{\alpha\beta.\gamma} \|$	$r_{\alpha\beta.\gamma}$	$\log\left(1-r_{\alpha\beta.\gamma}^2\right)^{\frac{1}{2}}$
$r_{12}=+0{,}80$	0,77815 − 1	+ 0,224	+ 0,576	0,76042 − 1	0,88043 − 1	0,87999 − 1	$r_{12.3}=+0{,}76$	0,81398 − 1
$r_{13}=-0{,}40$	0,96214 − 1	− 0,448	+ 0,048	0,68124 − 2	0,69644 − 1	0,98480 − 2	$r_{13.2}=+0{,}10$	0,99797 − 1
$r_{23}=-0{,}56$	0,91829 − 1	− 0,320	− 0,240	0,38021 − 1	0,74029 − 1	0,63992 − 1	$r_{23.1}=-0{,}44$	0,95411 − 1

mehr die Einhaltung eines gewissen Zwischenraums die Korrelation am stärksten zum Ausdruck bringt, weil manche Ursachen einer gewissen Zeit (Phasendifferenz) bedürfen, um den Höhepunkt ihrer Wirkung zu erreichen. Eine Welle in der Heiratsfrequenz zum Beispiel äußert ihre stärkste Wirkung auf die Geburtenhäufigkeit erst in etwas mehr als zwei Jahren.

So hat sich die Korrelationstheorie als ein brauchbares und wertvolles Mittel der Ursachenforschung erwiesen und besitzt in dieser Richtung fast unbegrenzte Anwendungsmöglichkeiten [1]).

97. Beispiel 1. In einigen Gebieten Englands sind aus einer zwanzigjährigen Periode die Ernten an Saaten (X_1) mit den Regenmengen (X_2) und den summierten Temperaturen über 42° F [2]) während des Frühlings (X_3) verglichen worden. Es ergaben sich folgende Zahlen [3]):

$$M_1 = 28{,}02 \text{ Zentner pro Acker}$$
$$M_2 = 4{,}91 \text{ Zoll}$$
$$M_3 = 594^0$$

$$\mu_1 = 4{,}42 \text{ Zentner}$$
$$\mu_2 = 1{,}10 \text{ Zoll}$$
$$\mu_3 = 85^0$$

$$r_{12} = +0{,}80$$
$$r_{13} = -0{,}40$$
$$r_{23} = -0{,}56.$$

Verlangt werden als letztes Ergebnis die Regressionsgleichungen.

Die zur Durchführung der Formelgruppe I nötige Rechnung ist nebenstehend in einer Tabelle zusammengestellt.

[1]) Die Korrelationstheorie ist auch auf psychophysische Vorgänge angewandt worden (vgl. Wilhelm Wirth, „Spezielle psychophysische Maßmethoden", Leipzig 1920).

[2]) Die Wahl dieser Grenztemperatur ist aus dem Gesichtspunkte erfolgt, daß tiefere Temperaturen auf das Wachstum noch keinen ersichtlichen Einfluß haben.

[3]) G. U. Yule, An Introduction to the Theory of Statistics. London 1932, S. 253.

Weiter folgen nun aus der Gruppe II folgende Werte für die mittleren Abweichungen 1. Ordnung:

$$\log \mu_{1.23} = 0{,}42154 \qquad \mu_{1.23} = 2.64$$
$$\log \mu_{2.13} = 0{,}77365 - 1 \qquad \mu_{2.13} = 0{,}59$$
$$\log \mu_{3.12} = 1{,}84567 \qquad \mu_{3.12} = 70{,}09$$

Zur Probe hat man beispielsweise die nachstehende zweifache Rechnung von $\log \mu_{1.23}$:

$$\begin{array}{rl} \log \mu_1 &= 0{,}64542 \\ \log\left(1-r_{12}^2\right)^{\frac{1}{2}} &= 0{,}77815 - 1 \\ \log\left(1-r_{13.2}^2\right)^{\frac{1}{2}} &= 0{,}99797 - 1 \\ \hline \log \mu_{1.23} &= 0{,}42154 \end{array} \qquad \begin{array}{rl} \log \mu_1 &= 0{,}64542 \\ \log\left(1-r_{13}^2\right)^{\frac{1}{2}} &= 0{,}96214 - 1 \\ \log\left(1-r_{12.3}^2\right)^{\frac{1}{2}} &= 0{,}81398 - 1 \\ \hline \log \mu_{1.23} &= 0{,}42154 \end{array}$$

Schließlich gibt die Ausrechnung der Formeln III:

$$\log b_{12.3} = 0{,}52788 \qquad b_{12.3} = 3{,}37$$
$$\log b_{13.2} = 0{,}56067 - 3 \qquad b_{13.2} = 0{,}0036$$
$$\log b_{21.3} = 0{,}23210 \qquad b_{21.3} = 1{,}71$$
$$\log |b_{23.1}| = 0{,}56790 - 3 \qquad b_{23.1} = -0{,}0037$$
$$\log b_{31.2} = 0{,}40893 \qquad b_{31.2} = 2{,}56$$
$$\log |b_{32.1}| = 1{,}71194 \qquad b_{32.1} = -51{,}52$$

Die Regressionsgleichungen, auf den Punkt $M_1 | M_2 | M_3$ bezogen, lauten demnach:

$$x_1 = 3{,}37\,x_2 + 0{,}0036\,x_3$$
$$x_2 = 1{,}71\,x_1 - 0{,}0037\,x_3$$
$$x_3 = 2{,}56\,x_1 - 51{,}52\,x_2,$$

und in den Variablen X_1, X_2, X_3 geschrieben:

$$X_1 = 9{,}33 + 3{,}37\,X_2 + 0{,}0036\,X_3$$
$$X_2 = -40{,}80 + 1{,}71\,X_1 - 0{,}0037\,X_3$$
$$X_3 = 775{,}23 + 2{,}56\,X_1 - 51{,}52\,X_2.$$

Zur Interpretation dieser Gleichungen diene das Folgende:

Zwischen Ernte und Regenmenge besteht eine ausgesprochene positive Korrelation ($r_{12.3} = +0{,}76$); der Koeffizient 3,37 weist auf eine deutliche Zunahme der Ernte mit wachsender Regenmenge hin.

Zwischen Ernte und Frühlingswärme besteht nur eine schwache positive Korrelation, wie dies sowohl der kleine Korrelationskoeffizient $r_{13.2} = +0{,}10$ wie auch der Regressionskoeffizient 0,0036 anzeigt.

Zwischen Regenmenge und Temperatur ist eine negative Korrelation vorhanden, wie an $r_{23.1} = -0{,}44$ und an den Koeffizienten $-0{,}0037$ und $-51{,}52$ zu erkennen ist, die besagen, daß mit zunehmender Temperatur die Regenmenge und umgekehrt mit zunehmender Regenmenge die Temperatur eine Abnahme erfährt.

Die Zusammenstellung der mittleren Abweichungen 0^ter und 2. Ordnung

$\mu_1 = 4{,}42$	$\mu_{1.23} = 2{,}64$
$\mu_2 = 1{,}10$	$\mu_{2.13} = 0{,}59$
$\mu_3 = 85$	$\mu_{3.12} = 70{,}09$

zeigt, daß durch Zusammenfassung aller drei Variablen durchwegs eine Herabsetzung platzgegriffen hat, was in dem Bau der Gleichungen (12) und (13) seine Begründung findet.

An sich wäre die zweite Gleichung, weil mit der kleinsten mittleren Abweichung behaftet, die verläßlichste; indessen entspricht die erste dem praktischen Bedürfnis am besten, denn auf die Ernte richtet sich vorwiegend das Interesse.

Zu bemerken ist noch, daß sich die totale negative Korrelation zwischen Ernte und Temperatur durch Zuziehung des dritten Faktors, der Regenmenge, in eine schwache positive verwandelt hat.

98. Beispiel 2. Im folgenden sollen die von Yule[1]) aufgefundenen Zusammenhänge, die sich auf das Fürsorgewesen beziehen, dargestellt werden. Als Variable seien eingeführt:

X_1: der prozentuale Anstieg der Zahl der Fürsorgeunterstützung Empfangenden vom Jahre 1881 bis zum Jahre 1891 in England;

X_2: das Gleiche bezüglich des Verhältnisses der die Unterstützung auswärts Empfangenden (offene Fürsorge) zu den in Anstalten Unterstützten (geschlossene Fürsorge);

X_3: das Gleiche bezüglich der Zahl der über 65 Jahre alten Personen an der Gesamtbevölkerung;

X_4: das Gleiche bezüglich der Bevölkerung selbst.

Aus den in 32 hauptstädtischen Bezirken erhobenen Zahlen ergaben sich folgende Grunddaten für die Lösung der Aufgabe:

		$r_{\alpha\beta}$	$\log\left(1 - r_{\alpha\beta}^2\right)$
$M_1 = 104{,}7$	$\mu_1 = 29{,}2$	$r_{12} = +0{,}52$	$0{,}93154 - 1$
$M_2 = 90{,}6$	$\mu_2 = 41{,}7$	$r_{13} = +0{,}41$	$0{,}96004 - 1$
$M_3 = 107{,}7$	$\mu_3 = 5{,}5$	$r_{14} = -0{,}14$	$0{,}99570 - 1$
$M_4 = 111{,}3$	$\mu_4 = 23{,}8$	$r_{23} = +0{,}49$	$0{,}94038 - 1$
		$r_{24} = +0{,}23$	$0{,}98820 - 1$
		$r_{34} = +0{,}25$	$0{,}98598 - 1$

Zum Verständnis der Zahlen sei erläuternd bemerkt:

Alle Daten für das Jahr 1881 werden gleich 100 gesetzt. Es haben also die Unterstützung Empfangenden in dem Dezennium um 4,7% zugenommen, das Verhältnis der außen Unterstützten zu den in Anstalten Untergebrachten hat sich um 9,4% seines Anfangswertes vermindert, die Zahl der über 65 Jahre alten Personen stieg um 7,7% und die Bevölkerung vermehrte sich um 11,3%. Den

[1]) G. U. Yule, An Introduction to the Theory of Statistics. London 1932, S. 241 u. f.

größten Schwankungen war X_2 unterworfen, den geringsten, wie vorauszusehen, die vom Altersaufbau stammende Zahl X_3. Die größte positive Korrelation zeigt der Anstieg der Unterstützten mit dem Verhältnis der außen und innen Unterstützten, die schwächste Korrelation tritt zwischen dem Anstieg der Unterstützten und dem der Bevölkerungszahl zutage.

Die Berechnung der Korrelationskoeffizienten 1. und 2. Ordnung geschieht am zweckmäßigsten nach jenen Gruppen, die in den Zählern auftreten; so bestehen die Zähler von

$$r_{12.3},\quad r_{13.2},\quad r_{23.1} \quad \text{aus } r_{12},\quad r_{13},\quad r_{23}$$

$$r_{12.4},\quad r_{14.2},\quad r_{24.1} \quad \text{aus } r_{12},\quad r_{14},\quad r_{24} \quad \text{u. s. w.}$$

die Zähler von

$$r_{12.34},\quad r_{13.24},\quad r_{23.14} \quad \text{aus } r_{12.3},\quad r_{14.3},\quad r_{24.3} \quad \text{u. s. w.}$$

In der zweifachen Berechnung der r zweiter Ordnung liegt eine Kontrolle der Rechnung.

Zur Bestimmung der Regressionskoeffizienten wäre die Kenntnis der mittleren Abweichungen 2. Ordnung erforderlich, indem z. B.

$$b_{12.34} = r_{12.34} \frac{\mu_{1.34}}{\mu_{2.34}}$$

ist; doch ist dies auch in mittleren Abweichungen 3. Ordnung ausdrückbar; denn

$$\mu_{1.234} = \mu_1 \left(1 - r_{14}^2\right)^{\frac{1}{2}} \left(1 - r_{13.4}^2\right)^{\frac{1}{2}} \left(1 - r_{12.34}^2\right)^{\frac{1}{2}}$$

$$\mu_{2.134} = \mu_2 \left(1 - r_{24}^2\right)^{\frac{1}{2}} \left(1 - r_{23.4}^2\right)^{\frac{1}{2}} \left(1 - r_{12.34}^2\right)^{\frac{1}{2}}$$

andererseits

$$\mu_{1.34} = \mu_1 \left(1 - r_{14}^2\right)^{\frac{1}{2}} \left(1 - r_{13.4}^2\right)^{\frac{1}{2}}$$

$$\mu_{2.34} = \mu_2 \left(1 - r_{24}^2\right)^{\frac{1}{2}} \left(1 - r_{23.4}^2\right)^{\frac{1}{2}}$$

daraus folgt aber

$$\frac{\mu_{1.34}}{\mu_{2.34}} = \frac{\mu_{1.234}}{\mu_{2.134}} = \frac{\mu_1 \left(1 - r_{14}^2\right)^{\frac{1}{2}} \left(1 - r_{13.4}^2\right)^{\frac{1}{2}}}{\mu_2 \left(1 - r_{24}^2\right)^{\frac{1}{2}} \left(1 - r_{23.4}^2\right)^{\frac{1}{2}}}$$

so daß auch

$$b_{12.34} = r_{12.34} \frac{\mu_{1.234}}{\mu_{2.134}}$$

ist und ähnlich die übrigen.

Was zur Berechnung der μ 3. Ordnung notwendig, ist bereits alles in den Tabellen Seite 220 enthalten, und man findet:

$$\begin{aligned} \log \mu_{1.234} &= 1{,}35740 \qquad & \mu_{1.234} &= 22{,}8 \\ \log \mu_{2.134} &= 1{,}50597 & \mu_{2.134} &= 32{,}1 \\ \log \mu_{3.124} &= 0{,}65773 & \mu_{3.124} &= 4{,}55 \\ \log \mu_{4.123} &= 1{,}32914 & \mu_{4.123} &= 21{,}3 \end{aligned}$$

Tab. 60 a. Berechnung der $r_{\alpha\beta.\gamma}$.

Korrelationskoeffizient (Nullter Ordnung)		Produktglied des Zählers	Zähler	Korrelationskoeffizient (Erster Ordnung)		log $\left(1-r_{\alpha\beta.\gamma}^2\right)^{\frac{1}{2}}$
12	+ 0,52	0,2009	0,3191	12.3	0,4013	0,96187 − 1
13	+ 0,41	0,2548	0,1552	13.2	0,2084	0,99035 − 1
23	+ 0,49	0,2132	0,2768	23.1	0,3553	0,97070 − 1
12	+ 0,52	− 0,0322	0,5522	12.4	0,5731	0,91355 − 1
14	− 0,14	0,1196	− 0,2596	14.2	− 0,3123	0,97772 − 1
24	+ 0,23	− 0,0728	0,3028	24.1	0,3580	0,97022 − 1
13	+ 0,41	− 0,0350	0,4450	13.4	0,4642	0,94730 − 1
14	− 0,14	0,1025	− 0,2425	14.3	− 0,2746	0,98297 − 1
34	+ 0,25	− 0,0574	0,3074	34.1	0,3404	0,97326 − 1
23	+ 0,49	0,0575	0,4325	23.4	0,4590	0,94863 − 1
24	+ 0,23	0,1225	0,1075	24.3	0,1274	0,99645 − 1
34	+ 0,25	0,1127	0,1373	34.2	0,1618	0,99425 − 1

Tab. 60 b. Berechnung der $r_{\alpha\beta.\gamma\delta}$.

Korrelationskoeffizient (Erster Ordnung)		Produktglied des Zählers	Zähler	Korrelationskoeffizient (Zweiter Ordnung)		log $\left(1-r_{\alpha\beta.\gamma\delta}^2\right)^{\frac{1}{2}}$
12.4	0,5731	0,2131	0,3600	12.34	0,457	0,94901 − 1
13.4	0,4642	0,2631	0,2011	13.24	0,276	0,98277 − 1
23.4	0,4590	0,2660	0,1930	23.14	0,266	0,98408 − 1
12.3	0,4013	− 0,0350	0,4363	12.34	0,457	0,94901 − 1
14.3	− 0,2746	0,0511	− 0,3257	14.23	− 0,359	0,97013 − 1
24.3	0,1274	− 0,1102	0,2376	24.13	0,270	0,98359 − 1
13.2	0,2084	− 0,0505	0,2589	13.24	0,276	0,98277 − 1
14.2	− 0,3123	0,0337	− 0,3460	14.23	− 0,359	0,97013 − 1
34.2	0,1618	− 0,0651	0,2269	34.12	0,244	0,98664 − 1
23.1	0,3553	0,1219	0,2334	23.14	0,266	0,98408 − 1
24.1	0,3580	0,1209	0,2371	24.13	0,270	0,98359 − 1
34.1	0,3404	0,1272	0,2132	34.12	0,244	0,98664 − 1

Jetzt kann an die Ausrechnung der 12 Regressionskoeffizienten geschritten werden; in der Regel wird man sich für die Wahl einer abhängigen Variablen entscheiden und braucht dann nur 3 Koeffizienten zu berechnen. Hier, wo es sich um die Illustration der Methode handelt, sollen alle ermittelt werden, um auch alle Regressionsgleichungen aufstellen zu können. Aus diesen sind die berechneten Werte der Koeffizienten zu ersehen. In der ersten Form — auf den „Punkt" $M_1 \mid M_2 \mid M_3 \mid M_4$ als Ursprung bezogen — lauten sie:

$$\begin{aligned}
x_1 &= 0{,}325\, x_2 + 1{,}383\, x_3 - 0{,}383\, x_4 \\
x_2 &= 0{,}644\, x_1 + 1{,}875\, x_3 + 0{,}405\, x_4 \\
x_3 &= 0{,}055\, x_1 + 0{,}038\, x_2 + 0{,}052\, x_4 \\
x_4 &= -0{,}336\, x_1 + 0{,}180\, x_2 + 1{,}146\, x_3
\end{aligned}$$

und in der zweiten Form

$$\begin{aligned}
X_1 &= -31{,}066 + 0{,}325\, X_2 + 1{,}383\, X_3 - 0{,}383\, X_4 \\
X_2 &= -223{,}841 + 0{,}644\, X_1 + 1{,}875\, X_3 + 0{,}405\, X_4 \\
X_3 &= 92{,}711 + 0{,}055\, X_1 + 0{,}038\, X_2 + 0{,}052\, X_4 \\
X_4 &= 6.747 - 0.336\, X_1 + 0{,}180\, X_2 + 1{,}146\, X_4 .
\end{aligned}$$

Unter den vier Gleichungen ist die dritte mit der kleinsten mittleren Abweichung behaftet. Vom praktischen Standpunkt aber ist die erste maßgebend, weil es sich um die Erforschung der Armutsverhältnisse handelt, soweit sie in der öffentlich-rechtlichen Fürsorgeunterstützung zum Ausdruck kommen.

Man kann den Sinn der ersten Gleichung dahin auslegen: Wenn sich die Verhältniszahlen X_2, X_3, X_4 je um 1 (ein Prozent) ändern, u. zw. zunehmen, so ändert sich X_1 im Durchschnitt um

$$0.325 + 1{,}383 - 0{,}383 = 1{,}325,$$

nimmt also um 1,325% zu.

Die Regressionskoeffizienten lassen das Maß des Einflusses der drei Faktoren erkennen: am stärksten macht sich der prozentuale Anteil der über 65 Jahre alten Personen geltend, am schwächsten das Verhältnis der auswärts Unterstützung Empfangenden zu den in Fürsorgeanstalten Unterstützten. Zunahme der Bevölkerung eines Fürsorgebezirkes geht mit einer prozentualen Abnahme der Unterstützten einher.

Die Zusammenstellung der gewöhnlichen Korrelationskoeffizienten mit den partiellen:

$$\begin{aligned}
r_{12} &= +0{,}52 & r_{12.34} &= +0{,}46 \\
r_{13} &= +0{,}41 & r_{13.24} &= +0{,}28 \\
r_{14} &= -0{,}14 & r_{14.23} &= -0{,}36 \\
r_{23} &= +0{,}49 & r_{23.14} &= +0{,}27 \\
r_{24} &= +0{,}23 & r_{24.13} &= +0{,}27 \\
r_{34} &= +0{,}25 & r_{34.12} &= +0{,}24
\end{aligned}$$

zeigt als auffälligste Erscheinung die Vergrößerung der negativen Korrelation zwischen dem prozentualen Anstieg der Zahl der Fürsorgeunterstützung Empfangenden und dem prozentualen Anstieg der Bevölkerungszahl[1]).

Weitere Beispiele. 1. R. Meerwarth[2]) verwendet zur Untersuchung der Frage, wodurch die hohen Geburtenziffern der einzelnen oberschlesischen Kreise hervorgerufen seien, die Korrelationsmethode. Er zieht als bestimmende Faktoren den relativen Anteil der katholischen Bevölkerung und den relativen Anteil der polnisch, sowie polnisch und deutsch sprechenden Bevölkerung in Betracht. Unter Zugrundelegung der Geburtenstatistik für das Jahr 1925 und der Volkszählung vom 16. Juni 1925 berechnet Meerwarth für die Beziehung zwischen Geburtenziffer und Anteil der katholischen Bevölkerung einen Korrelationskoeffizienten von + 0,547 und für die Beziehung zwischen Geburtenziffer und Anteil der polnisch, sowie polnisch und deutsch sprechenden Bevölkerung einen solchen von + 0,782. Der Zusammenhang zwischen Geburtenziffer und „polnischem Einschlag" ist somit wesentlich straffer als der Zusammenhang zwischen Geburtenziffer und Anteil der katholischen Bevölkerung. Weiter korreliert Meerwarth die statistischen Zahlen für den Anteil der katholischen Bevölkerung und für den polnischen Einschlag und erhält hierbei einen Korrelationskoeffizienten von + 0,591. Diese Berechnungsergebnisse weisen nach Meerwarth darauf hin, daß der entscheidende Einfluß auf die Geburtenhäufigkeit vom Einschlag der polnischen Bevölkerung herzukommen scheint; die Korrelation zwischen Geburtenziffer und Anteil der katholischen Be-

[1]) Man kann die für zwei Variable X_1, X_2 geltende Beziehung

$$\mu_{1.2}^2 = \mu_1^2 \left(1 - r_{12}^2\right)$$

auf den Fall einer mehrfachen Korrelation übertragen und als Maß der Gesamtkorrelation zwischen X_1 und $X_2, X_3, \ldots X_n$ eine Größe $R_{1.23\ldots n}$ durch folgende Gleichung definieren

$$\mu_{1.23\ldots n}^2 = \mu_1^2 \left(1 - R_{1.23\ldots n}^2\right).$$

Vergleicht man diese Gleichung mit (13), so ergibt sich der folgende Zusammenhang wischen der Größe R, die man als totalen Korrelationskoeffizienten bezeichnet, und den Korrelationskoeffizienten bis zur Ordnung $n - 2$:

$$1 - R_{1.23\ldots n}^2 = \left(1 - r_{12}^2\right)\left(1 - r_{13.2}^2\right)\left(1 - r_{14.23}^2\right)\ldots\left(1 - r_{1n.23\ldots(n-1)}^2\right),$$

aus dem hervorgeht, daß der totale Korrelationskoeffizient $R_{1.23\ldots n}$ größer ist als jeder der Korrelationskoeffizienten $r_{12}, r_{13.2} \ldots r_{1n.23\ldots(n-1)}$, insbesondere größer als r_{12}, und selbst wieder ein echter Bruch. Ähnliches gilt von $R_{2.134\ldots n}$ usw.

Bei dem ersten unserer Beispiele, betreffend die Abhängigkeit der Heuernte von Regenhöhe und Temperatur, ergeben sich

$$R_{1.23} = 0{,}80 \qquad R_{2.13} = 0{,}84 \qquad R_{3.12} = 0{,}57,$$

bei dem zweiten Beispiel

$$R_{1.234} = 0{,}63 \qquad R_{2.134} = 0{,}64 \qquad R_{3.124} = 0{,}56 \qquad R_{4.123} = 0{,}44;$$

ihre Vergleichung mit den entsprechenden partiellen und gewöhnlichen Korrelationskoeffizienten bestätigt die obigen Aussagen.

[2]) R. Meerwarth, Von dem Nutzen und den Grenzen der Statistik. Zeitschrift des Preußischen Statistischen Landesamts, 72. Jahrgang, S. 51 und 52.

völkerung wäre nur eine scheinbare. Meerwarth fügt den gewöhnlichen Korrelationskoeffizienten noch die partiellen hinzu. Der partielle Korrelationskoeffizient für den Zusammenhang zwischen Geburtenziffer und Anteil der katholischen Bevölkerung stellt sich bei rechnerischer Konstanthaltung der Maßzahl für den polnischen Einschlag auf + 0,169, während sich der partielle Korrelationskoeffizient für den Zusammenhang zwischen Geburtenziffer und polnischem Einschlag bei rechnerischer Konstanthaltung der Maßzahl für den katholischen Anteil auf + 0,680 beläuft. Diese beiden partiellen Korrelationskoeffizienten bestätigen den vorher festgestellten Zusammenhang.

2. C. Hempel[1]) untersucht die korrelative Abhängigkeit der ehelichen Fruchtbarkeitsziffer (β) von dem Heiratsalter der Frau (h) und der Ehedauer in vollendeten Jahren (d) unter Zugrundelegung einer besonderen Auszählung in Sachsen für die Jahre 1919 bis 1922. Hempel gewinnt für die vier Jahre die mittlere Regressionsgleichung

$$\beta = 28{,}544 - 0{,}855\, d - 1{,}298\, h$$

Hempel folgert hieraus, daß sich die eheliche Fruchtbarkeitsziffer um 1 vermindert, wenn man bei konstantem Heiratsalter der Frau in der Reihe der Ehejahre rechnerisch um 1,17 Jahre fortschreitet, oder wenn man bei konstanter Ehedauer das Heiratsalter der Frau um 0,77 Jahre erhöht. Das Heiratsalter der Frau ist somit von stärkerem Einfluß auf die Geburtenhäufigkeit als die Ehedauer.

3. Auf Grund der Volkszählung in Schottland im Jahre 1911 stellt J. C. Dunlop[2]) fest, daß zwischen der mittleren Zahl der Kinder einer Ehe (C), dem Heiratsalter der Frau (W), dem Heiratsalter des Mannes (H) und der Ehedauer (D) die folgende Regressionsgleichung

$$C = 3{,}299 - 0{,}076\, W - 0{,}024\, H - 0{,}267\, D$$

besteht. Aus dieser Regressionsgleichung folgt, daß im Durchschnitt die Pause zwischen zwei Geburten 3,7 Jahre beträgt und daß das Heiratsalter der Frau auf die Zahl der Kinder einer Ehe von größerem Einfluß ist als das Heiratsalter des Mannes.

4. Auf dem Gebiete der Zwillingsforschung ist neuerdings von K. Diehl und O. v. Verschuer[3]) die Korrelationsmethode angewandt worden. Hierbei handelt es sich teils um quantitative, teils um qualitative Merkmale. Es wurden Zwillingspaare hinsichtlich ihres Verhaltens gegenüber der Tuberkulose, hinsichtlich der Umweltverhältnisse, des Lebensalters und der Erbanlagen beobachtet. Das gewonnene Beobachtungsmaterial wurde in Gruppen gegliedert. Bezüglich des Verhaltens gegenüber der Tuberkulose werden die vier Gruppen C, c, d, D gebildet. Zur ersten Gruppe C (konkordant) rechnen Diehl und v. Verschuer diejenigen Zwillingspaare, die sich völlig gleich verhalten, zur zweiten Gruppe c (schwach konkordant) diejenigen, die geringe Unterschiede im Verhalten aufweisen, zur dritten Gruppe d (schwach diskordant) diejenigen, die sich in geringem Grade

[1]) C. Hempel, Statistische Untersuchungen über die eheliche Fruchtbarkeit. Dresden 1936, S. 87.

[2]) J. C. Dunlop, Report of the twelfth Decennial Census of Scotland. Bd. III, London 1913, S. XXXIX.

[3]) K. Diehl und O. v. Verschuer, Zwillingstuberkulose, Zwillingsforschung und erbliche Tuberkulosedisposition. Jena 1933, S. 430 u. f. Vgl. hierzu E. Weber, Variations- und Erblichkeitsstatistik. München 1935, J. F. Lehmanns Verlag. S. 133 u. f.

verschieden verhalten, und zur vierten Gruppe D (diskordant) diejenigen, die sich völlig verschieden verhalten. Bezüglich der Umweltverhältnisse werden die entsprechenden vier Umweltgruppen C, c, d, D unterschieden. Bezüglich des Lebensalters erfolgt die Gliederung nach Lebensjahrfünften und bezüglich der Erbanlagen nach eineiigen und zweieiigen gleichgeschlechtlichen Zwillingen. Diehl und v. Verschuer berechnen zunächst für je zwei Variable gewöhnliche Korrelationskoeffizienten und weiter zur schärferen Durchdringung der Zusammenhänge partielle Korrelationskoeffizienten 1. Ordnung. Für die Beziehung zwischen Tuberkuloseverhalten und Erbanlage stellt sich z. B. für die gleichgeschlechtlichen Zwillingspaare der gewöhnliche Korrelationskoeffizient auf + 0,47; die partiellen Korrelationskoeffizienten für diese Beziehung betragen unter Konstanthaltung der Umweltverhältnisse + 0,50 und unter Konstanthaltung des Lebensalters + 0,45. Die beiden partiellen Korrelationskoeffizienten weichen nur wenig von den gewöhnlichen Korrelationskoeffizienten ab. Hieraus ziehen Diehl und v. Verschuer die Schlußfolgerung, daß die Beziehung zwischen Tuberkuloseverhalten und Erbanlage weder durch Umwelt-, noch durch Altersbedingungen zu erklären ist. Für die Beziehung zwischen dem Tuberkuloseverhalten und der Umwelt ergibt sich für die eineiigen Zwillinge ein Korrelationskoeffizient von + 0,32 und für die zweieiigen gleichgeschlechtlichen ein solcher von + 0,56. Der partielle Korrelationskoeffizient berechnet sich für diese Beziehung unter Konstanthaltung des Lebensalters für die eineiigen Zwillinge auf + 0,34 und für die zweieiigen gleichgeschlechtlichen auf + 0,37. Durch Vergleichen der gewöhnlichen und partiellen Korrelationskoeffizienten gelangen Diehl und v. Verschuer zu der Schlußfolgerung, daß in der Beziehung des Tuberkuloseverhaltens zur Umwelt bei den zweieiigen gleichgeschlechtlichen Zwillingen ein schwach wirkender Alterseinfluß im Spiele zu sein scheint. Durch weitere Betrachtungen dieser Art kommen Diehl und v. Verschuer zu dem Ergebnis, daß die erbliche Veranlagung von maßgebender Bedeutung für die Entstehung und den Ablauf der Tuberkulose ist.

§ 9. Zerlegung von Zeitreihen.

99. Nach dem Vorgehen des „Harward University Committee on Economic Research“ zerlegt man die Schwankungen von Wirtschaftsreihen in vier Komponenten:

1. Trend (Grund- oder Hauptrichtung, Grund- oder Hauptverlauf, Grund- oder Hauptbewegung, Entwicklungstendenz),
2. Saisonschwankung (saisonmäßige oder jahreszeitliche Schwankungen),
3. Konjunkturschwankung (Konjunkturwellen, Konjunkturzyklen),
4. Sonstige Schwankungen (durch restliche Einflüsse hervorgerufene Schwankungen).

Die zur Herausarbeitung der ersten Komponente entwickelten Methoden wollen wir an dem Beispiel der deutschen Roheisengewinnung in den Jahren 1895 bis 1913 [1]) betrachten, das wir im wesentlichen in der anschaulichen Darstellungsweise von H. Hennig wiedergeben.

[1]) H. Hennig, Die Ausschaltung von saisonmäßigen und säkularen Schwankungen aus Wirtschaftskurven. Vierteljahrshefte zur Konjunkturforschung, 1. Jahrgang, 1926, Ergänzungsheft 1, S. 26 u. f.

Tab. 61. Die Roheisengewinnung in Deutschland 1895 bis 1913.

Jahr	Zeitwert x	Quadrat des Zeitwertes x^2	Jährliche Roheisen-gewinnung in 1000 t $10^{-3}s$	Produkt $10^{-3}s \cdot x$	Produkt $10^{-3}s \cdot x^2$	Trendwert $10^{-3}y$	Differenz $10^{-3}(s-y)$	Quadrat $10^{-6}(s-y)^2$	Trend-bereinigte Reihe $\frac{s}{y} \cdot 100$
1	2	3	4	5	6	7	8	9	10
1895	− 9	81	5 464	− 49 176	442 584	6 276	− 812	659 344	87,1
1896	− 8	64	6 373	− 50 984	407 872	6 484	− 111	12 321	98,3
1897	− 7	49	6 881	− 48 167	337 169	6 746	+ 135	18 225	102,0
1898	− 6	36	7 313	− 43 878	263 268	7 062	+ 251	63 001	103,6
1899	− 5	25	8 143	− 40 715	203 575	7 431	+ 712	506 944	109,6
1900	− 4	16	8 521	− 34 084	136 336	7 854	+ 667	444 889	108,5
1901	− 3	9	7 880	− 23 640	70 920	8 331	− 451	203 401	94,6
1902	− 2	4	8 530	− 17 060	34 120	8 862	− 332	110 224	96,3
1903	− 1	1	10 018	− 10 018	10 018	9 447	+ 571	326 041	106,0
1904	0	0	10 058	0	0	10 086	− 28	784	99,7
1905	+ 1	1	10 875	+ 10 875	10 875	10 778	+ 97	9 409	100,9
1906	+ 2	4	12 293	+ 24 586	49 172	11 524	+ 769	591 361	106,7
1907	+ 3	9	12 875	+ 38 625	115 875	12 324	+ 551	303 601	104,5
1908	+ 4	16	11 805	+ 47 220	188 880	13 178	− 1 373	1 885 129	89,6
1909	+ 5	25	12 645	+ 63 225	316 125	14 086	− 1 441	2 076 481	89,8
1910	+ 6	36	14 794	+ 88 764	532 584	15 048	− 254	64 516	98,3
1911	+ 7	49	15 574	+ 109 018	763 126	16 063	− 489	239 121	97,0
1912	+ 8	64	17 617	+ 140 936	1 127 488	17 132	+ 485	235 225	102,8
1913	+ 9	81	19 312	+ 173 808	1 564 272	18 255	+ 1 057	1 117 249	105,8
		570	206 971	+ 379 335	6 574 259	206 967		8 867 266	

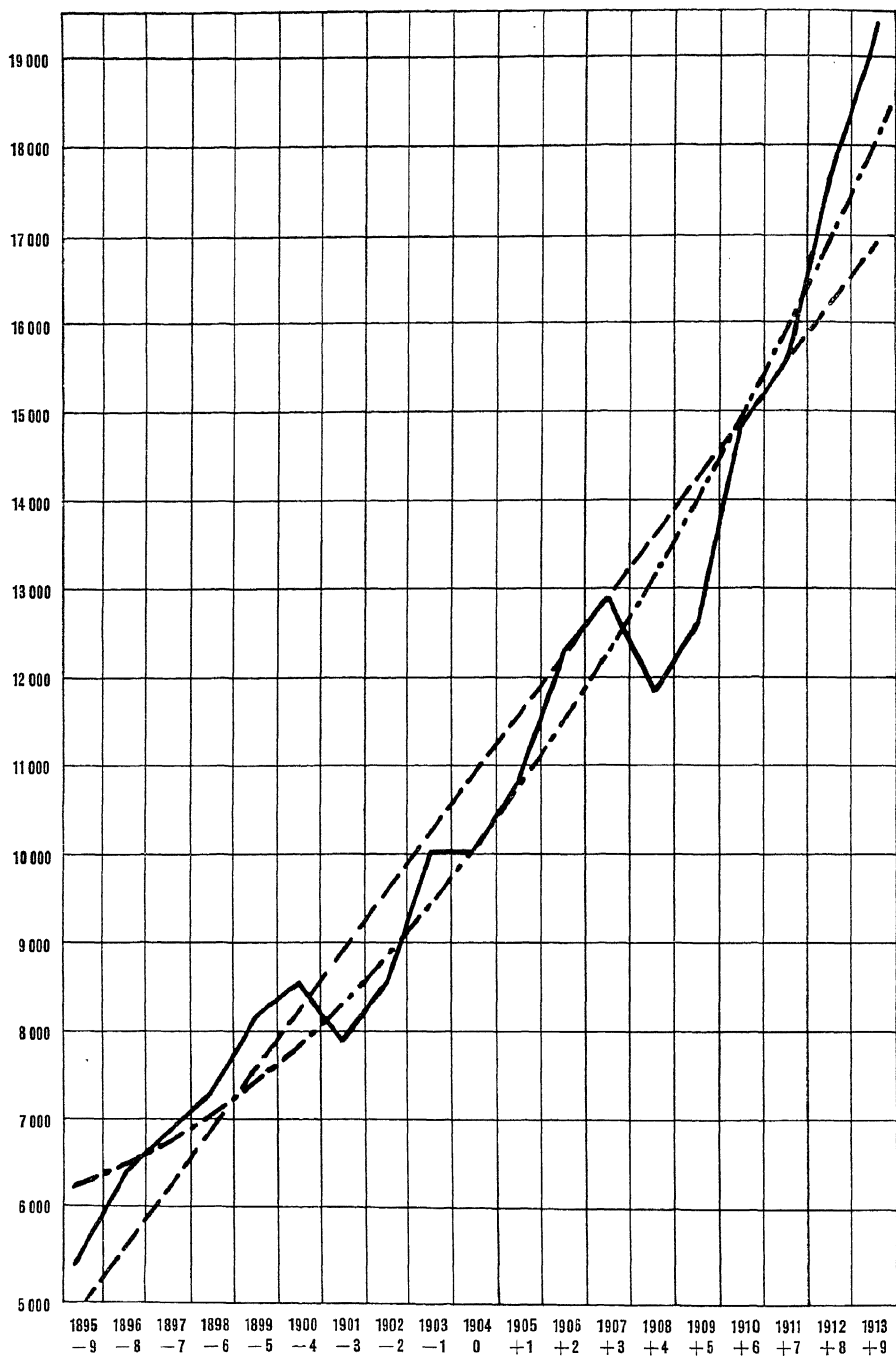



Fig. 30. Die Roheisengewinnung in Deutschland 1895—1913.

In Fig. 29 werden die in den Jahren 1895 bis 1913 gewonnenen Roheisenmengen (s) durch die ausgezogene unregelmäßige Kurve graphisch dargestellt. Die Jahresproduktionszahlen sind jeweils der Jahresmitte zugeordnet[1]).

In Fig. 29 sind weiter gestrichelte Kurven eingezeichnet worden, die die Grundbewegung, den Trend der deutschen Roheisengewinnung, veranschaulichen und als Trendlinien bezeichnet werden. Am einfachsten gewinnt man vom Trendverlauf eine Vorstellung, wenn man auf Grund der empirischen Zahlen das Punktbild herstellt und durch dieses nach Augenmaß eine glatte Linie hindurchlegt.

Für die zahlenmäßige Bestimmung der Trendlinien kommen zwei verschiedene Verfahren in Betracht: das Verfahren der gleitenden Durchschnitte und die mathematischen Methoden.

a) Verfahren der gleitenden Durchschnitte. Dieses Verfahren besteht darin, daß man zunächst aus den empirischen Zahlenwerten für die drei Jahre 1895, 1896 und 1897 das arithmetische Mittel bildet. Es sei bezeichnet mit M_{96}. Weiter bestimmt man das arithmetische Mittel für die Jahre 1896, 1897 und 1898. Dieses Mittel sei M_{97}. In dieser Weise fährt man fort. Es wird nun in dem Koordinatensystem für das Jahr 1896 der Wert M_{96} eingetragen, für das Jahr 1897 der Wert M_{97} u. s. f. Auf diese Weise erhält man einen Kurvenzug, der im allgemeinen nur noch geringe Unregelmäßigkeiten aufweist.

Ist dieser Kurvenzug bereits genügend glatt, dann läßt er sich als Trendlinie auffassen. Ist er dies noch nicht, dann kann man das angewendete Verfahren erweitern, indem man statt der drei Jahre jedesmal fünf Jahre zur Berechnung von Durchschnitten zusammenfaßt.

Ist auch die auf diesen fünfjährigen Durchschnitten beruhende Kurve noch nicht genügend glatt, dann kann man das Verfahren auf einen noch größeren Zeitraum anwenden, etwa auf 6, 7 u. s. f. Jahre. Jedoch ist zu beachten, daß bei Zusammenfassung einer größeren Zahl von Jahren die charakteristischen Verhältnisse in den einzelnen Jahren mehr oder weniger verloren gehen. Im allgemeinen möchte man nicht mehr als zehn Jahre zusammenfassen. Bei einer geraden Zahl von Jahren nimmt man von den beiden Flügeljahren die Hälfte[2]).

b) Mathematische Methoden. Zur Herausarbeitung der Grundbewegung kann man grundsätzlich alle Funktionsarten verwenden. In der praktischen statistischen Forschung werden hauptsächlich ganze rationale Funktionen und Exponentialfunktionen angesetzt.

In Fig. 29 sind zwei Trendlinien eingezeichnet, eine gerade und eine parabolische. Wir wollen zunächst die gerade Trendlinie betrachten, sie ist gestrichelt.

---

[1]) Bei der graphischen Darstellung von Zeitreihen bedient man sich mitunter des logarithmischen Maßstabs. Man trägt nicht die Werte selbst, sondern ihre Logarithmen als Ordination auf. Gleichen Strecken auf der Ordinatenachse entsprechen dann nicht gleichgroße absolute Veränderungen, sondern gleichgroße relative Veränderungen der Werte. Die Anwendung des logarithmischen Maßstabs empfiehlt sich dann, wenn die Unterschiede der kleinen Werte der Reihe recht scharf herausgearbeit werden sollen und wenn verschiedene Zahlenreihen, deren Werte verschiedenen Größenordnungen angehören, in einer Zeichnung dargestellt werden sollen. Vgl. hierzu z. B. Wirtschaft und Statistik 1937, S. 69, 83, 145, 622. Vgl. auch W. Schweer, Über graphische Methoden in der Versicherungs-Mathematik. Archiv für mathematische Wirtschafts- und Sozialforschung, 1936, Bd. II, S. 98 u. f.

[2]) Die Nachteile der gleitenden Durchschnitte hat R. Wagenführ in seiner Schrift „Statistik leicht gemacht“ (Hamburg 1934, S. 82) dargelegt.

Die ihr entsprechende ganze rationale Funktion vom ersten Grade setzen wir in der Form

$$y = ax + b \tag{1}$$

an. Die Variable x durchläuft die Zeitwerte. Man setzt zweckmäßig für das in der Mitte des Zeitraums liegende Jahr 1904 den Zeitwert 0 an und demzufolge für das Jahr 1895 den Zeitwert -9, für 1896 den Zeitwert -8 u. s. f., für 1913 den Zeitwert $+9$. Die Variable y gibt den Trendwert an, der zum Zeitwert x gehört. Die Konstante b ist gleich dem Trendwert zur Zeit $x = 0$, und die Konstante a bedeutet das Richtverhältnis der Geraden und kennzeichnet somit die Steigung der Trendgeraden.

Die Bestimmung der Konstanten a und b erfolgt mittels der Methode der kleinsten Quadrate. Man bestimmt die Abweichung der empirisch-statistischen Zahlen für die gewonnenen Roheisenmengen von den nach der Gleichung (1) für den betreffenden Zeitwert sich ergebenden Trendwert. Es seien bezeichnet die empirisch-statistischen Werte der Reihe nach mit $s_1, s_2, \ldots,$ die Trendwerte mit $y_1, y_2, \ldots$ und die Zeitwerte mit $x_1, x_2, \ldots$ Hierbei beziehen sich in unserem Beispiel die Werte s_1, y_1 auf den Zeitwert $x_1 = -9$ (d. h. auf das Jahr 1895). Die Abweichungen der empirisch-statistischen Werte von den entsprechenden Trendwerten lassen sich somit in der Form darstellen:

$$s_1 - y_1 = s_1 - ax_1 - b$$
$$s_2 - y_2 = s_2 - ax_2 - b \text{ usw.}$$

Nach der Methode der kleinsten Quadrate werden die Abweichungen $s_i - y_i$ quadriert, die erhaltenen Quadrate addiert und die beiden Konstanten a und b so bestimmt, daß die Summe der Quadrate möglichst klein wird. Dies führt zu folgendem Ansatz für n Zeitwerte:

$$\sum_1^n (s_i - ax_i - b)^2 = \text{Minimum.}$$

Durch partielle Differentiation nach a und b erhält man:

$$-2 \sum_1^n (s_i - ax_i - b)\, x_i = 0$$

$$-2 \sum_1^n (s_i - ax_i - b) = 0.$$

Hieraus ergeben sich die Normalgleichungen:

$$a \sum_1^n x_i^2 + b \sum_1^n x_i = \sum_1^n s_i x_i$$

$$a \sum_1^n x_i + nb = \sum_1^n s_i.$$

Infolge der Wahl der Zeitwerte ist

$$\sum x_i = 0.$$

Die Determinante der Koeffizienten verschwindet nicht. Man erhält:

$$a = \frac{\begin{vmatrix} \Sigma s x & 0 \\ \Sigma s & n \end{vmatrix}}{\begin{vmatrix} \Sigma x^2 & 0 \\ 0 & n \end{vmatrix}} = \frac{\Sigma s x}{\Sigma x^2}$$

$$b = \frac{\begin{vmatrix} \Sigma x^2 & \Sigma s x \\ 0 & \Sigma s \end{vmatrix}}{\begin{vmatrix} \Sigma x^2 & 0 \\ 0 & n \end{vmatrix}} = \frac{\Sigma s}{n}.$$

In der Tabelle 61 wird die Berechnung durchgeführt. Sie liefert:

$$a = \quad 665\,500$$
$$b = 10\,893\,000.$$

Die Gleichung der Trendlinie lautet somit:

$$y = 665\,500\,x + 10\,893\,000.$$

In Fig. 29 ist außer der geraden Trendlinie noch eine parabolische Trendlinie (Parabel zweiter Ordnung —.—.—.) eingezeichnet. Ihr entspricht die Gleichung:

$$y = a x^2 + b x + c. \tag{2}$$

Die Bestimmung der Konstanten erfolgt mittels des folgenden Ansatzes:

$$\sum_1^n (s_i - a x_i^2 - b x_i - c)^2 = \text{Minimum}.$$

Durch partielle Differentiation nach a, b und c erhält man:

$$-2 \sum_1^n (s_i - a x_i^2 - b x_i - c)\, x_i^2 = 0$$

$$-2 \sum_1^n (s_i - a x_i^2 - b x_i - c)\, x_i = 0$$

$$-2 \sum_1^n (s_i - a x_i^2 - b x_i - c) \quad = 0.$$

Hieraus folgen, da

$$\sum_1^n x_i^3 = 0, \quad \sum_1^n x_i = 0,$$

die Normalgleichungen:

$$a \sum_1^n x_i^4 + c \sum_1^n x_i^2 = \sum_1^n s_i x_i^2$$

$$b \sum_1^n x_i^2 = \sum_1^n s_i x_i$$

$$a \sum_1^n x_i^2 + n c = \sum_1^n s_i .$$

Somit ergibt sich:

$$a = \frac{\begin{vmatrix} \Sigma s x^2 & 0 & \Sigma x^2 \\ \Sigma s x & \Sigma x^2 & 0 \\ \Sigma s & 0 & n \end{vmatrix}}{\begin{vmatrix} \Sigma x^4 & 0 & \Sigma x^2 \\ 0 & \Sigma x^2 & 0 \\ \Sigma x^2 & 0 & n \end{vmatrix}} = \frac{n \Sigma s x^2 - \Sigma s \Sigma x^2}{n \Sigma x^4 - (\Sigma x^2)^2}$$

$$b = \frac{\begin{vmatrix} \Sigma x^4 & \Sigma s x^2 & \Sigma x^2 \\ 0 & \Sigma s x & 0 \\ \Sigma x^2 & \Sigma s & n \end{vmatrix}}{\begin{vmatrix} \Sigma x^4 & 0 & \Sigma x^2 \\ 0 & \Sigma x^2 & 0 \\ \Sigma x^2 & 0 & n \end{vmatrix}} = \frac{\Sigma s x}{\Sigma x^2}$$

$$c = \frac{\begin{vmatrix} \Sigma x^4 & 0 & \Sigma s x^2 \\ 0 & \Sigma x^2 & \Sigma s x \\ \Sigma x^2 & 0 & \Sigma s \end{vmatrix}}{\begin{vmatrix} \Sigma x^4 & 0 & \Sigma x^2 \\ 0 & \Sigma x^2 & 0 \\ \Sigma x^2 & 0 & n \end{vmatrix}} = \frac{\Sigma x^4 \Sigma s - \Sigma x^2 \Sigma s x^2}{n \Sigma x^4 - (\Sigma x^2)^2}.$$

Die Berechnung von a, b und c erfolgt nach den in der Tab. 61 zusammengestellten Zahlenreihen und unter Beachtung der folgenden Formeln:

$$2 \sum_{1}^{\nu} x^2 = 2[1^2 + 2^2 + 3^2 + \cdots + \nu^2] = \tfrac{2}{6} \nu (\nu + 1)(2\nu + 1)$$

$$2 \sum_{1}^{\nu} x^4 = 2[1^4 + 2^4 + 3^4 + \cdots + \nu^4] = \tfrac{2}{30} \nu (\nu + 1)(2\nu + 1)(3\nu^2 + 3\nu - 1).$$

Hierbei ist $\nu = \frac{n-1}{2} = 9.$

Man erhält:

$$a = 26\,915$$
$$b = 665\,500$$
$$c = 10\,085\,760$$

und für die Trendlinie:

$$y = 26\,915 x^2 - 665\,500 x - 10\,085\,760.$$

In Tab. 61 sind diese Trendwerte in Spalte 7 aufgeschrieben.

Es gilt, nun noch die Standardabweichung vom Trend festzustellen, die den Grad der Anpassung oder das Maß für das Genügen der Trendlinie kennzeichnet. Zu diesem Zwecke berechnen wir in Spalte 8 die Abweichungen der empirischen Werte von den entsprechenden Trendwerten $(s - y)$ und bilden das quadratische Mittel dieser Abweichungen.

Standardabweichung vom Trend $= \sqrt{\frac{\Sigma (s-y)^2}{}}$. Die Berechnung ergibt für die Trendgerade 992000 (= 9,1 % vom Mittel der s-Werte) und für die parabolische Trendlinie (Parabel 2. Ordnung) 683000 (= 6,3 % vom Mittel der s-Werte)[1].

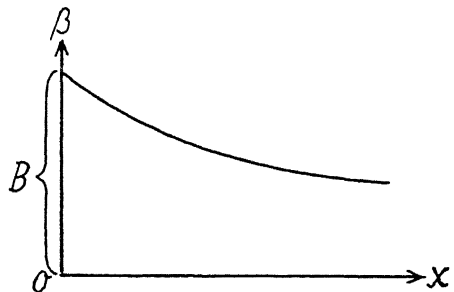
Die Ausschaltung der Grundbewegung erfolgt in der Weise, daß man die einzelnen Trendwerte gleich einer Konstanten, z. B. gleich 1 oder gleich 100 setzt und die zugehörigen empirischen Zahlen dementsprechend umrechnet. Die vom Trend „bereinigte" empirisch-statistische Zahl s' gewinnt man somit nach der Formel

$$s' = \frac{s}{y} \cdot 100.$$

Die Standardabweichung kann vermindert und damit der Grad der Anpassung der Trendlinie an das Beobachtungsmaterial erhöht werden, wenn man als Trendlinien Parabeln 3. und höherer Ordnung wählt. Das Berechnungsverfahren ist das gleiche wie bei Parabeln 2. Ordnung, nur wird die Berechnung etwas umfangreicher. Im allgemeinen kommt man in der praktischen Forschung mit Parabeln 2. Ordnung aus.

Anschließend sei bemerkt, daß eine Scheidung zwischen „Trend" und „langer Welle" vorzunehmen ist. Nach den Darlegungen von E. Wagemann[2]) und R. Wagenführ[3]) kennzeichnet der Trend die Entwicklungstendenz während eines langen Zeitraumes, z. B. während eines Jahrhunderts. Eine lange Welle dagegen erstreckt sich nur auf einen Zeitraum von etwa 25 bis 30 Jahren.

Fig. 31.

100. Die Trendlinien haben auch in der bevölkerungsstatistischen Forschung eine Bedeutung. Die rückläufige Bewegung der Totgeborenenquote β im Zeitraum 1851 bis 1913 erfolgte näherungsweise nach der in Fig. 30 gezeichneten Kurve.

[1]) H. Peter führt in seiner Schrift „Statistik und Theorie in den Wirtschaftswissenschaften" (Stuttgart 1935, S. 85) aus, daß sich die gleitenden Durchschnitte enger an die Ursprungskurve anschmiegen als eine nach der Methode der kleinsten Quadrate angenäherte Kurve. Der Grund hierfür liegt darin, daß die gleitenden Durchschnitte aus einer verhältnismäßig kleinen Zahl von Ursprungswerten abgeleitet werden, während bei der Methode der kleinsten Quadrate sämtliche Werte der Ursprungskurve zugleich angesetzt werden. Vgl. hierzu die Ausführungen von H. v. Stackelberg in dem Aufsatz „Die grundlegenden Hypothesen der neueren Preisanalyse". Archiv für mathematische Wirtschafts- und Sozialforschung, Bd. 1, 1935, S. 84 u. f.

[2]) E. Wagemann, Struktur und Rhythmus der Weltwirtschaft. Stuttgart, Leipzig 1931, S. 144.

[3]) R. Wagenführ, Die Bedeutung des Außenmarktes für die deutsche Industriewirtschaft. Sonderheft 41 des Instituts für Konjunkturforschung 1936. S. 38 u. f.

Diese Kurve wollen wir mittels der e-Funktion in der Form darstellen:

$$\beta = B e^{-h'x} \tag{3}$$

Hierbei bezeichnet x in laufender Folge die einzelnen Kalenderjahre des Beobachtungszeitraums, B die Totgeborenenquote im Anfangsjahr $x = 0$ [1]) und h' den Grad des relativen Rückgangs der Totgeborenenquote. Für h' ergibt sich sofort aus Gleichung (3) durch Differenzieren nach x:

$$\frac{1}{\beta}\frac{d\beta}{dx} = -h'. \tag{4}$$

Die Größe h' wird wie die entsprechenden Konstanten in der praktischen statistischen Forschung positiv angesetzt. Die Kurve der Fig. 30 entspricht der Gleichung (3); denn die erste Ableitung ist negativ und die zweite positiv. Schreiben wir die Gleichung (3) getrennt für Knaben und Mädchen, indem wir mit β_m bzw. β_w die Totgeborenenquote der Knaben, bzw. die der Mädchen bezeichnen, so erhalten wir:

$$\beta_m = B_m e^{-h_1 x}$$
$$\beta_w = B_w e^{-h_2 x}.$$

Aus diesen beiden Gleichungen gewinnen wir durch Differentiation nach x und anschließende Division

$$\frac{d\beta_m}{d\beta_w} = h\frac{\beta_m}{\beta_w}. \tag{5}$$

Hierbei ist $h = \frac{h_1}{h_2}$. Durch Integration folgt aus (5)

$$\frac{\beta_m}{\beta_w^h} = C. \tag{6}$$

Die beiden Konstanten C und h bestimmen wir durch logarithmischen Ausgleich mittels der Methode der kleinsten Quadrate, indem wir zunächst (6) in der logarithmischen Form $\lg\beta_m = h\lg\beta_w + \lg C$ schreiben und sodann die Bedingung stellen:

$$\Sigma(-\lg\beta_m + h\lg\beta_w + \lg C)^2 = [(-\lg\beta_m + h\lg\beta_w + \lg C)^2] = \text{Minimum.}^{2)}$$

Wir erhalten:

$$h = \frac{n\,[\lg\beta_m \lg\beta_w] - [\lg\beta_m]\,[\lg\beta_w]}{n\,[(\lg\beta_w)^2] - [\lg\beta_w]^2}, \tag{7 a}$$

$$\lg C = \frac{[\lg\beta_m]\,[(\lg\beta_w)^2] - [\lg\beta_m \lg\beta_w]\,[\lg\beta_w}{n\,[(\lg\beta_w)^2] - [\lg\beta_w]^2} \tag{7 b}$$

[1]) Wir ordnen hier aus sachlich-anschaulichen Gründen dem Anfangsjahr und nicht dem in der Mitte des Zeitraumes liegenden Kalenderjahr die Abszisse $x=0$ zu.

[2]) Wir verwenden hier, um die Ausdrücke für h und $\lg C$ einfacher schreiben zu können, an Stelle des Zeichens Σ die in der Ausgleichungsrechnung übliche Summenbezeichnung [ ].

Hierbei bedeutet n die Anzahl der Kalenderjahre, über die sich die Untersuchung der Bewegung der Totgeborenenquote erstreckt. Führt man die Berechnung für die verschiedenen europäischen Länder durch, so erhält man im allgemeinen

$$h > 1.$$

Nur in Sachsen und in Belgien liegt die Beziehung vor

$$h < 1,$$

d. h. mit Ausnahme von Sachsen und Belgien war der relative Rückgang der Totgeborenenquote beim männlichen Geschlecht größer als beim weiblichen.

Im Zeitraum 1914 bis 1926 ist im allgemeinen die Totgeborenenquote gestiegen. Diese ansteigende Bewegung läßt sich durch die Trendgleichung

$$\beta = B e^{k'' x} \tag{8}$$

darstellen. Mittels dieser Gleichung lassen sich dieselben Umformungen und Berechnungen vornehmen wie mit der Gleichung (3). Im besonderen findet man, daß im allgemeinen gilt

$$h > 1.$$

Zur Nachprüfung der Anpassung der von uns gewählten Trendlinie an das Beobachtungsmaterial bestimmen wir die mittlere Abweichung der empirischen Totgeborenenquoten von den Trendwerten. Wir stellen hierbei fest, daß die mittlere Abweichung in den einzelnen europäischen Ländern nicht mehr als 5 % der Konstanten C beträgt. Für das Deutsche Reich stellt sich diese Prozentziffer für den Zeitraum 1851 bis 1913 auf 3,5. Im allgemeinen weisen die Trendlinien in der Bevölkerungsstatistik einen höheren Grad der Anpassung an das Beobachtungsmaterial auf als in der Wirtschaftsstatistik.

Für die Werte der Sterblichkeit im ersten Lebensjahr eignet sich als Trendlinie am besten der Ellipsenbogen (Fig. 31).

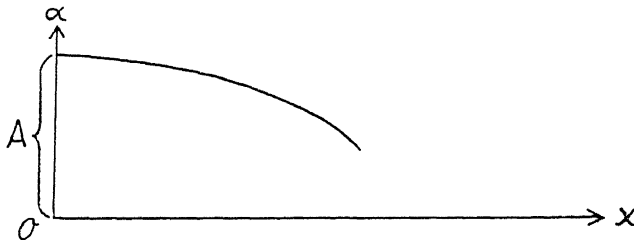



Fig. 32.

Wir setzen zunächst die Gleichung eines Kegelschnitts in der Form an

$$x^2 (1 - \varepsilon^2) + \alpha^2 = \frac{p^2}{1 - \varepsilon^2}. \tag{9}$$

Hierbei bedeuten ε die numerische Exzentrizität und p den Hauptparameter. Für ε erhält man auf Grund der statistischen Zahlen für die europäischen Länder nach der Methode der kleinsten Quadrate

$$\varepsilon < 1,$$

d. h. das betrachtete Kurvenstück ist ein Teil einer Ellipse. Es sei von vornherein, um falschen Deutungen vorzubeugen, darauf hingewiesen, daß man die Trendlinie nur für den Beobachtungszeitraum zeichnen kann und nicht darüber hinaus fortsetzen darf. Denn es ist ohne weiteres klar, daß es sinnlos wäre, die Trendlinie bis zum Schnitt mit der x-Achse zu verfolgen. Durch Differentiation der Gleichung (9) nach x erhält man

$$\frac{d\alpha}{dx} = \frac{-x(1-\varepsilon^2)}{\alpha}.$$

Dividiert man die letzte Gleichung durch α und setzt für x^2 den Ausdruck aus (9) ein, so gelangt man zu

$$\frac{d\alpha}{\alpha} = -\frac{x(1-\varepsilon^2)^2\,dx}{p^2 - x^2(1-\varepsilon^2)^2}.$$

Das im Nenner in additiver Verbindung auftretende relativ kleine Glied $x^2(1-\varepsilon^2)^2$ kann vernachlässigt werden.

Schreibt man die letzte Gleichung getrennt für die beiden Geschlechter auf und dividiert man die eine auf diese Weise entstehende Gleichung durch die andere, so findet man:

$$\frac{d\alpha_m}{\alpha_m} = k\,\frac{d\alpha_w}{\alpha_w}.$$

Durch Integration der letzten Gleichung ergibt sich:

$$\frac{\alpha_m}{\alpha_w^k} = c. \tag{10}$$

Die beiden Konstanten k und c können mittels der Methode der kleinsten Quadrate nach den Gleichungen (7a) und (7b) bestimmt werden. Man erhält durchgängig:

$$k < 1.$$

Hieraus folgt, daß der Rückgang der Säuglingssterblichkeit beim weiblichen Geschlecht relativ größer war als beim männlichen.

Spaltet man die Gleichung (6) nach der Legitimität auf, so erhält man die beiden neuen Gleichungen

$$\frac{\beta_{me}}{\beta_{ve}^{\eta}} = C_e$$

$$\frac{\beta_{mu}}{\beta_{wu}^{i}} = C_u,$$

wobei sich die Indizes e bzw. u auf die Ehelichen bzw. Unehelichen beziehen. Nach dem statistischen Zahlenmaterial ist im allgemeinen

$$C_e > C_u.$$

Hieraus folgt die Ungleichung:

$$\frac{\beta_{wu}^{\,i}}{\beta_{we}^{\,\eta}} > \frac{\beta_{mu}}{\beta_{me}} .$$

Mittels der Ausgleichungsrechnung unter Verwendung der empirischen Zahlen gilt in der Regel:

$$\eta < i, \quad \eta > 1, \quad i > 1.$$

Also ist

$$\frac{\beta_{wu}}{\beta_{we}} > \frac{\beta_{wu}^{\,i}}{\beta_{we}^{\,\eta}} .$$

Kombiniert man die beiden letzten Ungleichungen, so findet man

$$\frac{\beta_{wu}}{\beta_{we}} > \frac{\beta_{mu}}{\beta_{me}} . \tag{11}$$

Diese Ungleichung läßt sich folgendermaßen in Worten ausdrücken: Die Übertotgeburtlichkeit der unehelichen Kinder ist beim weiblichen Geschlecht größer als beim männlichen.

Die Gleichung (10) läßt sich ebenso wie die Gleichung (6) nach der Legitimität aufteilen. Man erhält:

$$\frac{\alpha_{me}}{\alpha_{we}^{\,\varkappa}} = c_e$$

$$\frac{\alpha_{mu}}{\alpha_{wu}^{\,l}} = c_u .$$

Im allgemeinen ist nach den statistischen Zahlen

$$c_e > c_u ,$$

folglich

$$\frac{\alpha_{wu}^{\,l}}{\alpha_{we}^{\,\varkappa}} > \frac{\alpha_{mu}}{\alpha_{me}}$$

Weiter findet man mittels der Ausgleichungsrechnung auf Grund des statistischen Zahlenmaterials

$$l > \varkappa, \quad l < 1, \quad \varkappa < 1.$$

Hieraus ergibt sich:

$$\frac{\alpha_{wu}}{\alpha_{we}} > \frac{\alpha_{wu}^{\,l}}{\alpha_{we}^{\,\varkappa}}$$

und durch Kombination der beiden letzten Ungleichungen folgt:

$$\frac{\alpha_{wu}}{\alpha_{we}} > \frac{\alpha_{mu}}{\alpha_{me}} . \tag{12}$$

Diese letzte Ungleichung besagt, daß die Übersterblichkeit der Unehelichen beim weiblichen Geschlecht größer ist als beim männlichen.

Die beiden Ungleichungen (11) und (12) lassen sich von einem gemeinsamen Standpunkt aus betrachten. In Art. 56 ergab sich die These, daß das weibliche Geschlecht von den äußeren Verhältnissen stärker abhängig ist als das männliche und daß infolgedessen das weibliche Geschlecht auf die äußeren Verhältnisse in höherem Grade reagiert als das männliche. Von dieser These aus gewinnt man sofort eine Erklärung für die beiden Ungleichungen (11) und (12). Infolge des stärkeren Reagierens des weiblichen Geschlechts auf äußere Verhältnisse leiden die Mädchen mehr unter den Unbilden der Illegitimität als die Knaben, und zwar gilt dies sowohl für das vorgeburtliche als auch für das nachgeburtliche Lebensstadium. Die Ungleichung (12) ist bereits in Kap. 56 auf empirischem Wege hergeleitet worden.

Durch Verbindung des statistischen und des deduktiven Forschungsverfahrens läßt sich die Ungleichung (12), die wir zuerst auf empirischem Wege und im vorstehenden mittels der Trendmethode gefunden haben, auch mittels der Korrelationsmethode gewinnen. Nach den Forschungen G. v. Mayr's[1]) besteht zwischen der Säuglingssterblichkeit α und der Übersterblichkeit der Knaben $\frac{\alpha_m}{\alpha_w}$ sowie zwischen der Säuglingssterblichkeit α und der Übersterblichkeit der Unehelichen $\frac{\alpha_u}{\alpha_e}$ eine negativ-korrelative Beziehung. Spaltet man die erste Korrelation, die in der Form $\alpha \Big| \frac{\alpha_m}{\alpha_w}$ geschrieben sei, nach der Legitimität in die beiden Korrelationen $\alpha_e \Big| \frac{\alpha_{me}}{\alpha_{we}}$ und $\alpha_u \Big| \frac{\alpha_{mu}}{\alpha_{wu}}$ auf, so gelangt man auf deduktivem Wege, da $\alpha_e < \alpha_u$ ist, zu der Ungleichung

$$\frac{\alpha_{me}}{\alpha_{we}} > \frac{\alpha_{mu}}{\alpha_{wu}}. \qquad (12\,a)$$

Die Schlußfolgerung ist um so zwingender, je größer der Korrelationskoeffizient (absolut) ist. Ist dieser 1, dann gilt die Schlußfolgerung vollkommen exakt. Die für die verschiedenen Länder und für die verschiedenen Zeiträume berechneten Koeffizienten der Korrelation $\alpha \Big| \frac{\alpha_m}{\alpha_w}$ schwanken um den Mittelwert von $-0{,}6$. Spaltet man die zweite Korrelation $\alpha \Big| \frac{\alpha_u}{\alpha_e}$ nach dem Geschlecht in die beiden Korrelationen $\alpha_m \Big| \frac{\alpha_{mu}}{\alpha_{me}}$ und $\alpha_w \Big| \frac{\alpha_{wu}}{\alpha_{we}}$ auf, so findet man durch analoge Überlegung, da $\alpha_m > \alpha_w$ ist, die Ungleichung

$$\frac{\alpha_{mu}}{\alpha_{me}} < \frac{\alpha_{wu}}{\alpha_{we}}. \qquad (12\,b)$$

Die beiden Ungleichungen (12 a) und (12 b) lassen sich durch eine einfache Rechenoperation ineinander überführen und stimmen mit der Ungleichung (12) vollkommen überein.

[1]) G. v. Mayr, Die Sterblichkeit der Kinder während des ersten Lebensjahres in Süddeutschland, insbesondere in Bayern. Zeitschrift des Königl.-Bayerischen Statistischen Bureau 1870, S. 212, und Statistik und Gesellschaftslehre, 2. Bd. Bevölkerungsstatistik, Tübingen 1926, S. 456 u. f.

§ 9. Zerlegung von Zeitreihen.

Anschließend sei noch die Darstellung der rückläufigen Bewegung der Säuglingssterblichkeit mittels der e-Funktion behandelt. Wir wollen zunächst folgende e-Funktion ansetzen:

$$\alpha = 2A - Ae^{\mu' x}.$$

Spaltet man diese Ungleichung nach dem Geschlecht und der Legitimität auf, so erhält man nach analogen Umformungen wie früher:

$$\frac{d\alpha_{m}}{d\alpha_{we}} = \mu \cdot \frac{2A_{me} - \alpha_{me}}{2A_{we} - \alpha_{we}}$$

Durch Integration erhält man für die Ehelichen:

$$\frac{2A_{me} - \alpha_{me}}{(2A_{we} - \alpha_{we})^{\mu}} = C_e$$

und für die Unehelichen:

$$\frac{2A_{mu} - \alpha_{mu}}{(2A_{wu} - \alpha_{wu})^{\nu}} = C_u.$$

Mittels der Methode der kleinsten Quadrate stellt man auf Grund der deutschen Statistik von 1901 bis 1934 fest, daß

$$C_e > C_u$$
$$\mu < \nu, \quad \mu < 1, \quad \nu < 1.$$

Auf diese Weise erhält man schließlich folgende Ungleichung:

$$\frac{A_{wu} - \frac{\alpha_{wu}}{2}}{A_{we} - \frac{\alpha_{we}}{2}} > \frac{A_{mu} - \frac{\alpha_{mu}}{2}}{A_{me} - \frac{\alpha_{me}}{2}}$$

Diese Ungleichung, die ohne jede Ausnahme im Zeitraum 1901 bis 1934 für die deutsche Statistik gilt, kann in folgender Weise gedeutet werden: $A - \frac{\alpha}{2}$ stellt den bis zur Hälfte der jetzigen Sterblichkeit in die Zukunft vorgeschobenen Sterblichkeitsrückgang dar. Der so konstruierte Übersterblichkeitsrückgang der Unehelichen ist somit beim weiblichen Geschlecht größer als beim männlichen.

Weiter sei zur Darstellung der Bewegung der Säuglingssterblichkeit die folgende Trendfunktion angesetzt:

$$\alpha = A + 1 - e^{\varrho x}.$$

Auf Grund dieser Trendfunktion erhält man die folgende Ungleichung:

$$\frac{A_{wu} + 1 - \alpha_{wu}}{A_{we} + 1 - \alpha_{u}} > \frac{A_{mu} + 1 - \alpha_{mu}}{A_{me} + 1 - \alpha_{n}}.$$

Der Sinn dieser Ungleichung, die allerdings einige Ausnahmen aufweist, geht dahin, daß nach ihr der um 1 vergrößerte Übersterblichkeitsrückgang der Unehelichen beim weiblichen Geschlecht größer ist als beim männlichen[1]).

101. Zur Herausarbeitung der Saisonschwankungen ist von dem amerikanischen Konjunkturforscher Persons eine Methode entwickelt worden, die man als Gliedziffernmethode bezeichnet. Diese Methode wollen wir an dem Beispiel der gesamten arbeitstäglichen Wagengestellung der Preußisch-hessischen und Oldenburgischen Eisenbahnen von Dezember 1904 bis Dezember 1913[2]) behandeln. Die empirischen Zahlen für diese Betrachtung sind in Tab. 62a zusammengestellt.

Die Grundzahl eines Monats sei allgemein mit s_ν und die des Vormonats mit $s_{\nu-1}$ bezeichnet. Wir bilden das Verhältnis:

$$g_\nu = \frac{s_\nu}{s_{\nu-1}}$$

und bezeichnen g_ν als die Gliedziffer des ν-ten Monats. Die aus Tab. 62a sich ergebenden Gliedziffern werden in der folgenden Tab. 62b zusammengestellt. Für jede Spalte der Tab. 62b wird das ungewogene arithmetische Mittel berechnet. Diese Mittelwerte seien bezeichnet mit $\bar{g}_1, \bar{g}_2, \ldots \bar{g}_{12}$. Bei der Bestimmung dieser Mittelwerte werden diejenigen Gliedziffern ausgelassen, die stark aus dem Rahmen herausfallen. Um diese letzteren Gliedziffern zu ermitteln, stellt man die einzelnen Gliedziffern graphisch dar. Dies geschieht in Tab. 62c, in der die Gruppen gebildet werden: 84 bis 85, 85 bis 86, . . ., 111 bis 112. Die Einordnung der einzelnen Gliedziffern in die Gruppen erfolgt auf Grund von Tab. 62b unter Berücksichtigung der ersten Dezimalen. Die untere Gruppengrenze wird zur Gruppe gerechnet, die obere dagegen nicht.

Aus den Mittelwerten $\bar{g}_1, \bar{g}_2, \ldots \bar{g}_{12}$ berechnet man die Kettenziffern k_1, $k_2, \ldots k_{12}$.

$$k_1 = \bar{g}_1$$
$$k_2 = \bar{g}_1 \cdot \bar{g}_2$$
$$\vdots$$
$$k_{12} = \bar{g}_1 \cdot \bar{g}_2 \ldots \bar{g}_{12}.$$

In der Regel weicht k_{12} von 1 ab. Dies ist eine Folge der konjunkturellen und der unregelmäßigen Einflüsse. Es sei:

$$k_{12} - 1 = \delta.$$

[1]) Vgl. F. Burkhardt, Über die Verbindung des deduktiven und des statistischen Forschungsverfahrens mittels mathematischer Denkformen, dargestellt an der Statistik der vor- und nachgeburtlichen Sterblichkeit. Archiv für mathematische Wirtschafts- und Sozialforschung, Bd. II, Heft 3, 1936, S. 149 u. f.

[2]) Vgl. P. Lorenz, Kurzer Abriß zur Methodik der Kurvenbehandlung in E. Wagemann, Konjunkturlehre. Berlin 1928, S. 236 u. f.

Tab. 62a. Zahl der gestellten Wagen pro Arbeitstag (in 1000).

Jahr	Januar	Februar	März	April	Mai	Juni	Juli	August	Sept.	Okt.	Nov.	Dez.
1904	.	.	.	.	.	.	.	.	.	.	.	89,5
1905	77,8	86,4	95,7	96,7	97,1	95,2	95,2	97,8	102,5	109,3	112,4	106,9
1906	98,6	101,4	105,4	105,5	106,1	101,6	105,3	107,3	109,8	116,0	117,4	108,0
1907	100,3	101,4	112,8	113,7	109,8	112,4	112,0	113,9	115,9	124,6	127,8	117,5
1908	103,0	111,9	114,1	115,0	116,1	109,9	114,9	115,9	120,5	127,9	127,9	108,6
1909	103,7	106,8	113,6	118,5	121,5	117,0	119,9	121,4	125,2	136,7	136,6	119,1
1910	113,6	117,0	126,2	123,5	124,5	121,8	125,3	130,4	133,7	146,6	147,8	130,9
1911	118,0	126,2	132,5	136,4	138,6	130,7	134,9	138,0	142,1	148,0	150,9	140,4
1912	129,3	139,7	145,5	146,9	149,0	144,8	146,3	151,6	159,6	162,6	164,2	161,8
1913	147,5	156,2	162,3	156,8	155,3	159,4	158,0	161,4	167,7	176,1	178,6	160,5

Tab. 62b. Berechnung der Gliedziffern (in %).

Jahr	Januar / Dez.	Februar / Januar	März / Februar	April / März	Mai / April	Juni / Mai	Juli / Juni	August / Juli	Sept / August	Okt. / Sept.	Nov. / Okt.	Dez. / Nov.
1905	[86,9]	[111,1]	[110,8]	101,0	100,4	98,0	100,0	102,7	104,8	106,6	102,8	95,1
1906	92,2	102,9	103,9	100,1	100,6	95,8	103,6	101,9	102,3	105,6	101,2	92,0
1907	92,9	101.1	[111,2]	100,8	[96,6]	[102,4]	99,6	101,7	101,8	107,5	102,6	91,9
1908	[87,7]	108,6	102,0	100,8	101,0	94,7	104,5	100,0	104,0	106,1	100,9	[84,9]
1909	[95,5]	103,0	106,4	104,3	102,5	96,3	102,5	101,3	103,1	109.2	99,9	87,2
1910	[95,4]	103,0	107,9	[97,9]	100,8	97,8	103,0	104,1	102,5	109,6	100,8	88,6
1911	90,1	106,9	105,0	102,9	101,6	94,3	103,2	102,3	103,0	104,2	102,0	93,0
1912	92,1	108,0	104,2	101,0	101,4	97,2	101,0	103,6	105,3	[101,9]	101,0	[98,5]
1913	91,2	105,8	103,9	[96,6]	99,0	[102,6]	99,1	102,2	103,9	105,0	101,4	89,9

Die Differenz δ teilen wir entweder arithmetisch oder geometrisch auf die 12 Monate auf. Die korrigierten Kettenziffern stellen sich bei der ersten Aufteilung auf:

$$k'_1 = k_1 - \frac{\delta}{12}$$

$$k'_2 = k_2 - 2\frac{\delta}{12}$$

$$\vdots$$

$$k'_{12} = k_{12} - 12\frac{\delta}{12} = 1$$

und bei der zweiten Aufteilung auf:

$$k''_1 = \frac{k_1}{(1+\delta)^{\frac{1}{12}}}$$

$$k''_2 = \frac{k_2}{(1+\delta)^{\frac{2}{12}}}$$

$$\vdots$$

$$k''_{12} = \frac{k_{12}}{1+\delta} = 1.$$

Aus den korrigierten Kettenziffern berechnet man sodann das ungewogene arithmetische Mittel:

$$\bar{k}' = \frac{k'_1 + k'_2 + \cdots + k'_{12}}{12}$$

bzw.

$$\bar{k}'' = \frac{k''_1 + k''_2 + \cdots + k''_{12}}{12}.$$

Auf diese letzteren Mittelwerte bezieht man die korrigierten Kettenziffern und erhält so die Saisonindexziffern $I'_1, I'_2, \ldots I'_{12}$ bzw. $I''_1, I''_2, \ldots I''_{12}$.

$$I'_1 = \frac{k'_1}{\bar{k}'}, \quad I''_1 = \frac{k''_1}{\bar{k}''}$$

$$I'_2 = \frac{k'_2}{\bar{k}'}, \quad I''_2 = \frac{k''_2}{\bar{k}''}$$

$$\vdots \qquad \vdots$$

$$I'_{12} = \frac{k'_{12}}{\bar{k}'}, \quad I''_{12} = \frac{k''_{12}}{\bar{k}''}.$$

In der folgenden Tab. 62d werden für unser Beispiel die Werte $\bar{g}$, k, k', k'', I', I'' und weiter die saisonbereinigten Werte für 1913 zusammengestellt.

Die Ausschaltung der Saisonschwankungen geschieht in der Weise, daß man die empirischen Zahlen (s) zu den entsprechenden Saisonindexziffern ins Verhältnis setzt:

$$\text{saisonbereinigter Wert} = \frac{\text{empirischer Wert}}{\text{Saisonindexziffer}} = \frac{s}{I}.$$

Tab. 62 d. Zusammenstellung.

Monat	100 $\bar{y}$	100 k	100 k'	100 k''	100 I'	100 I''	$\frac{s}{I'} \cdot 10^{-3}$	$\frac{s}{I''} \cdot 10^{-3}$
Januar	91,7	91,7	91,2	91,3	90,8	90,9	162,4	162,3
Februar	104,9	96,2	95,3	95,3	94,9	94,9	164,6	164,6
März	104,8	100,8	99,4	99,5	99,0	99,1	163,9	163,8
April	101,6	102,4	100,6	100,6	100,2	100,2	156,5	156,5
Mai	100,9	103,3	101,0	101,0	100,6	100,6	154,4	154,4
Juni	96,3	99,5	96,7	96,9	96,3	96,5	165,5	165,2
Juli	101,8	101,3	98,1	98,2	97,7	97,8	161,7	161,6
August	102,2	103,5	99,8	99,9	99,4	99,5	162,4	162,2
September	103,4	107,0	102,9	102,8	102,4	102,4	163,8	163,8
Oktober	106,7	114,2	109,6	109,2	109,1	108,7	161,4	162,0
November	101,4	115,8	110,7	110,3	110,2	109,8	162,1	162,7
Dezember	91,1	105,5	100,0	100,0	99,6	99,6	161,1	161,1

Zur Bestimmung der Saisonschwankungen ist die Gliedziffernmethode auch in der amtlichen Statistik angewandt worden. O. Donner hat eingehend die Saisonschwankungen der wichtigsten Wirtschaftsvorgänge in Deutschland untersucht. Die Ergebnisse dieser Untersuchungen hat Donner niedergelegt in den Sonderheften 6 und 11 des Instituts für Konjunkturforschung und in seinem Buche „Statistik" (Hamburg 1937, S. 72 u. f.).

Weiter hat G. Gräbner die Saisonschwankungen im Außenhandel einer eingehenden Untersuchung unterzogen. Gräbner unterscheidet Form- und Weitenschwankungen von Außenhandelsreihen. Das Nähere hierüber führt Gräbner in dem Aufsatz „Saisonschwankungen im Außenhandel" im Deutschen Statistischen Zentralblatt 1935, Sp. 101 u. f. aus.

Die Ergebnisse der Ausschaltung von Saisonschwankungen im Textilaußenhandel, im Außenhandel mit Nahrungs- und Genußmitteln und im Außenhandel mit verschiedenen Produktionsmitteln und Verbrauchsgütern sind vom Statistischen Reichsamt in den Vierteljahrsheften zur Statistik des Deutschen Reichs, Jahrgang 1934, Heft 4, S. 76; Jahrgang 1935, Heft 1, S. 159; Jahrgang 1936, Heft 1, S. 60 veröffentlicht worden.

Die Saisonschwankungen der Exportquote hat R. Wagenführ eingehend behandelt. Die Ergebnisse dieser Untersuchung hat Wagenführ in Sonderheft 41 des Instituts für Konjunkturforschung, S. 23 u. f. veröffentlicht.

Die mathematischen Methoden für die Zerlegung von Zeitreihen spielen auch in der betriebswirtschaftlichen Statistik eine Rolle, so z. B. bei der rechnerischen Isolierung des Saison- und Konjunkturgewinns. Es sei in diesem Zusammenhang

auf die grundlegenden Untersuchungen von A. Hoffmann[1]) hingewiesen. Hoffmann bestimmt zunächst die Trendkurve (Gerade und Parabel zweiter Ordnung) der Umsätze. Die Abweichungen der beobachteten Werte von den berechneten Werten werden in Prozent der letzteren ausgedrückt. Aus den prozentualen Abweichungen gleichnamiger Monate werden die arithmetischen Mittel gebildet, die die Saisonschwankung charakterisieren. Durch Subtraktion dieser Saisonschwankungszahlen von den prozentualen Abweichungen ergeben sich die reinen Konjunkturschwankungen. Bei diesem Verfahren gelingt es, die Zerlegung in einem Zuge durchzuführen.

Zwischen der volkswirtschaftlichen und betriebswirtschaftlichen Statistik bestehen eine Reihe von Beziehungen. Diese hat W. Morgenroth[2]) vom Standpunkt der volkswirtschaftlichen Statistik aus in richtungsweisenden Abhandlungen untersucht. Im betriebswirtschaftlichen Geschehen zeigen sich auch im Laufe des Tages Schwankungen, die in Parallele zu den Saisonschwankungen gestellt werden können; solche Schwankungen treten z. B. im Personenverkehr der Reichsbahn zutage Hierüber hat H. Kellerer[3]) eingehende statistische Untersuchungen angestellt. Aus einer sogenannten Verkehrsuhr kann die Zahl der in den einzelnen Tagesstunden ankommenden Reisenden abgelesen werden mit Unterscheidung von Benützern von Monats-, Wochen- und Einzelkarten. Kellerer stellt auf Grund einer Statistik für Hamburg fest, daß die Hauptverkehrsstunde des Gesamtverkehrs und für die Benützer von Monatskarten 8 bis 9, bei Wochenkarten 7 bis 8 und bei Einzelkarten 19 bis 20 Uhr ist. Zur schärferen Kennzeichnung dieser Schwankungen am Tage dient die Maßzahl der „Konzentration". Man berechnet sie durch Division der Reisendenzahl in der Hauptverkehrsstunde durch den durchschnittlichen Stundenverkehr. Sie stellt sich nach der Hamburger Statistik für den Verkehr mit Monatskarten auf 3,6; Wochenkarten auf 2,4 und Einzelkarten auf 1,6.

Die trend- und saisonbereinigten Werte stellen die Konjunkturschwankungen (einschl. sonstigen Schwankungen) dar.

§ 10. Die Anwendung der Methode der kleinsten Quadrate auf geldliche Ausgleichsprobleme in der Verwaltung.

102. Bei geldlichen Verteilungen in der Verwaltung handelt es sich zunächst um die Wahl der Verteilungsmaßstäbe. Beim interkommunalen Finanzausgleich, bei dem ein bestimmter Geldbetrag auf die Gemeinden eines Landes verteilt werden soll, kommen als Verteilungsmaßstäbe in Betracht: Steueraufkommen (a) und Bevölkerungszahl (b); beim Lastenausgleich: Gesamtlast (l), Steuereingang (s) und Bevölkerungszahl (b); bei der Verteilung von Mitteln für den Wohnungsbau: fehlende Wohnungen und Bevölkerungszahl und andere.

[1]) A. Hoffmann, Der Gewinn der kaufmännischen Unternehmung. Leipzig 1929, S. 255 u. f.

[2]) W. Morgenroth, „Betriebswirtschaftliche und allgemeine Statistik." Allgemeines Statistisches Archiv, 19. Bd., 1929, S. 334 u. f. und „Zusammenarbeit der volkswirtschaftlichen und betriebswirtschaftlichen Statistik". Allgemeines Statistisches Archiv, 20. Bd., 1930, S. 350 u. f.

[3]) H. Kellerer, Verkehrsstatistik. Eine vergleichende Gesamtdarstellung der Ziele und Lösungswege. Berlin 1936, S. 222 u. f.

Stehen mehrere Verteilungsmaßstäbe zur Wahl, so liefert die Korrelationsmethode Anhaltspunkte für die Auswahl. Man berechnet zu diesem Zwecke Korrelationskoeffizienten unter Zugrundelegung der regionalen Werte der Verteilungsmaßstäbe. Liegt für zwei Verteilungsmaßstäbe der Korrelationskoeffizient nahe an $+1$, dann wirken beide Verteilungsmaßstäbe in gleicher Weise für eine Gruppe von Gebietskörperschaften günstig und für die andere Gruppe ungünstig. Da es nicht erwünscht ist, nur solche Verteilungsmaßstäbe zu verwenden, die in gleicher Weise einseitig wirken, so wird man im allgemeinen solche Verteilungsmaßstäbe auswählen, die gegenseitig keinen starken positiv-korrelativen Zusammenhang aufweisen. Im Gegenteil, man wird die Maßstäbe bevorzugen, die in negativer Korrelation zueinander stehen.

Die ausgewählten Verteilungsmaßstäbe (Verteilungsmaßzahlen) seien allgemein mit a, b, c, . . . bezeichnet. Die Verteilungsrechnung wird in der Weise durchgeführt, daß man aus den Verteilungsmaßzahlen Schlüsselzahlen berechnet. In der Verwaltungspraxis nimmt man die Berechnung der Schlüsselzahl entweder durch additive oder multiplikative Verknüpfung der Verteilungsfaktoren vor. Stehen die Verteilungsmaßstäbe im direkten Verhältnis zum Anteilsbetrag, so setzt man in der Regel die Schlüsselzahlen (s_1, s_2, . . .) in folgender Form[1]) an:

$$s_1 = \alpha \bar{a}_1 + \beta \bar{b}_1 + \gamma \bar{c}_1 + \cdots$$
$$s_2 = \alpha \bar{a}_2 + \beta \bar{b}_2 + \gamma \bar{c}_2 + \cdots$$

Hierbei sind $\bar{a}_1, \bar{a}_2, \ldots, \bar{b}_1, \bar{b}_2, \ldots, \bar{c}_1, \bar{c}_2, \ldots$ die normierten Werte der Ursprungszahlen $a_1, a_2, \ldots, b_1, b_2, \ldots, c_1, c_2, \ldots$ Die Normierung nimmt man meistens in der Weise vor, daß man die Summenzahlen

$$A = a_1 + a_2 + \cdots$$
$$B = b_1 + b_2 + \cdots$$

gleich 1 000 000 setzt und die einzelnen Summanden dementsprechend umrechnet. Weiter sind α β, γ . . . positive rationale Zahlen, deren Summe gleich 1 ist; sie bedeuten die statistischen Gewichte, mit denen die Verteilungsmaßstäbe in die Rechnung eingeführt werden. Ist z. B. $\alpha = 0{,}6$, $\beta = 0{,}3$, $\gamma = 0{,}1$, so heißt das, daß der Verteilungsmaßstab a zu 60 %, der Verteilungsmaßstab b zu 30 % und der Verteilungsmaßstab c zu 10 % berücksichtigt werden.

Bei multiplikativer Verknüpfung der Verteilungsfaktoren erhält man die folgende allgemeine Schlüsselzahl für den Fall, daß alle Verteilungsmaßstäbe in direktem Verhältnis zum Anteilsbetrag stehen:

$$s = a^{\alpha}\, b^{\beta}\, c^{\gamma}$$

[1]) Die Indizes 1, 2 . . . beziehen sich der Reihe nach auf die erste, zweite . . . Gebietskörperschaft (z. B. Gemeinde).

die sich durch Logarithmieren auf die Gestalt bringen läßt:

$$\lg s = \alpha \lg a + \beta \lg b + \gamma \lg c.$$

Bei multiplikativer Verknüpfung haben die Normierung der Verteilungsmaßzahlen und die Bedingung $\alpha + \beta + \ldots = 1$ keine besondere Bedeutung.

Stehen einzelne Verteilungsmaßstäbe im indirekten Verhältnis zum Anteilsbetrag, so werden diese bei additiver Verknüpfung mit negativem Vorzeichen und bei multiplikativer Verknüpfung mit negativem Exponenten in die Formel für die Schlüsselzahl eingesetzt. Steht z. B. unter den Verteilungsmaßstäben $a, b, c \ldots$ der Verteilungsmaßstab b in indirektem Verhältnis zum Anteilsbetrag, so ergibt sich als allgemeine Schlüsselzahl bei additiver Verknüpfung

$$s = \alpha \bar{a} - \beta \bar{b} + \gamma \bar{c} + \cdots$$

und bei multiplikativer Verknüpfung

$$s = \frac{a^\alpha}{b^\beta} c^\gamma \ldots$$

Im allgemeinen empfiehlt sich die additive Verknüpfung dann, wenn alle Verteilungsmaßstäbe im direkten Verhältnis zum Anteilsbetrag stehen und die multiplikative Verknüpfung dann, wenn einzelne Verteilungsmaßstäbe im indirekten Verhältnis zum Anteilsbetrag stehen. In den anderen beiden Fällen ergeben sich in der Regel zu starke Schwankungen der Schlüsselzahlen, die vom Standpunkt der praktischen Verwaltung aus nicht erwünscht sind.

Für die zahlenmäßige Bestimmung der statistischen Gewichte legen wir das Prinzip der kleinsten Kürzungen oder der geringsten Härte zugrunde. Wir wollen diese Betrachtungen für den interkommunalen Finanzausgleich mit den beiden Verteilungsfaktoren: Steueraufkommen (a) und Bevölkerungszahl (b) durchführen. Es gibt Gemeinden, die die Verteilung lediglich nach dem Steueraufkommen wünschen. Es sind das die Gemeinden, für die

$$\bar{a} > \bar{b}.$$

Andererseits gibt es Gemeinden, die als Verteilungsmaßstab lediglich die Bevölkerungszahl haben möchten. Für diese Gemeinden ist

$$\bar{a} < \bar{b}.$$

Für die erste Gruppe von Gemeinden ist:

$$\frac{a}{A} \cdot 1\,000\,000 > \frac{b}{B} \cdot 1\,000\,000$$

$$\frac{a}{b} > \frac{A}{B},$$

d. h. der Kopfsatz des Aufkommens in diesen Gemeinden liegt über dem Landesdurchschnitt.

Für die zweite Gruppe von Gemeinden ist:

$$\frac{a}{b} < \frac{A}{B},$$

d. h. der Kopfsatz des Aufkommens liegt unter dem Landesdurchschnitt.

Bei voller Berücksichtigung der Interessen der Gemeinden der ersten Gruppe (I) bzw. der zweiten Gruppe (II)[1]) würde die Schlüsselzahl lauten

$$s_{\text{I}} = \bar{a}_{\text{I}}$$

bzw.

$$s_{\text{II}} = \bar{a}_{\text{II}}\,.$$

Die mittlere Schlüsselzahl setzen wir in der Form

$$s = \alpha\,\bar{a} + \beta\,\bar{b}$$

oder zum leichteren Verständnis in der Form

$$s = x\,\bar{a} + (1-x)\,\bar{b}$$

an. Die Kürzung, die das maximale Interesse einer Gemeinde der ersten Gruppe erfährt, wenn man von der maximalen Schlüsselzahl, die sich bei voller Berücksichtigung der Interessen ergibt, zur mittleren Schlüsselzahl übergeht, ist

$$K_{\text{I}} = \bar{a}_{\text{I}} - [x\,\bar{a}_{\text{I}} + (1-x)\,\bar{b}_{\text{I}}] = (1-x)\,(\bar{a}_{\text{I}} - \bar{b}_{\text{I}}).$$

Entsprechend erhalten wir für die Kürzung bei den Gemeinden der zweiten Gruppe

$$K_{\text{II}} = \bar{b}_{\text{II}} - [x\,\bar{a}_{\text{II}} + (1-x)\,\bar{b}_{\text{II}}] = x\,(\bar{b}_{\text{II}} - \bar{a}_{\text{II}}).$$

Wir führen nun das Prinzip der geringsten Härte mittels der Methode der kleinsten Quadrate durch:

$$\Sigma\,K_{\text{I}}^2 + \Sigma\,K_{\text{II}}^2 = \text{Minimum}.$$

Wir erhalten durch Differenzieren

$$-(1-x)\,\Sigma\,(\bar{a}_{\text{I}} - \bar{b}_{\text{I}})^2 + x\,\Sigma\,(\bar{b}_{\text{II}} - \bar{a}_{\text{II}})^2 = 0.$$

Hieraus ergibt sich

$$x = \frac{\Sigma\,(\bar{a}_{\text{I}} - \bar{b}_{\text{I}})^2}{\Sigma\,(\bar{a}_{\text{I}} - \bar{b}_{\text{I}})^2 + \Sigma\,(\bar{b}_{\text{II}} - \bar{a}_{\text{II}})^2}\,.$$

Die numerische Durchführung der Rechnung ergab für die sächsischen Amtshauptmannschaften Werte, die schwanken um einen Mittelwert von

$$1925/26 \qquad x = 0{,}6$$
$$1930/31 \qquad x = 0{,}5,$$

d. h. es war nach den Verhältnissen von 1925/26 die Verteilung dann am gerechtesten, wenn 60 % nach dem Aufkommen und 40 % nach der Bevölkerungszahl verteilt wurden. Nach den Verhältnissen von 1930/31 war das gerechteste Verhältnis 50 : 50.

Diese Verschiedenheit im Ergebnis hängt damit zusammen, daß eine Abhängigkeit von der Konjunktur besteht. Diese Abhängigkeit hat folgenden Ursprung. Es ist

$$\Sigma\,\bar{a}_{\text{I}} + \Sigma\,\bar{a}_{\text{II}} = \Sigma\,\bar{b}_{\text{I}} + \Sigma\,\bar{b}_{\text{II}} \text{ oder}$$
$$\Sigma\,\bar{a}_{\text{I}} - \Sigma\,\bar{b}_{\text{I}} = \Sigma\,\bar{b}_{\text{II}} - \Sigma\,\bar{a}_{\text{II}} \text{ oder}$$
$$\Sigma\,(\bar{a}_{\text{I}} - \bar{b}_{\text{I}}) = \Sigma\,(\bar{b}_{\text{II}} - \bar{a}_{\text{II}}).$$

Der Ausdruck $\Sigma\,(\bar{a}_{\text{I}} - \bar{b}_{\text{I}})^2$ ist umso größer, je größer die einzelnen Werte der Differenz $(\bar{a}_{\text{I}} - \bar{b}_{\text{I}})$ sind.

---

[1]) Dies würde dadurch herbeigeführt werden, daß für die Gemeinden der ersten Gruppe (I) $\alpha = 1$ und $\beta = 0$ und für die Gemeinden der zweiten Gruppe (II) $\alpha = 0$ und $\beta = 1$ gesetzt wird.

In wirtschaftlich günstigen Zeiten haben die Industriegemeinden ein hohes Steueraufkommen. Die Differenz $(\bar{a}_I - \bar{b}_I)$ ist groß. Die statistische Gewichtszahl x nimmt also einen hohen Wert an. In wirtschaftlich ungünstigen Zeiten vermindert sich das Steueraufkommen in den Großstädten und Industriegemeinden, die in der Hauptsache zur ersten Gruppe (I) gehören, in besonderem Grade. Demzufolge vermindern sich auch die Differenzen $(\bar{a}_I - \bar{b}_I)$. Infolgedessen stellt sich der Wert für x in wirtschaftlich ungünstigen Zeiten etwas niedriger als in wirtschaftlich günstigen[1]).

Zur weiteren Erläuterung und Veranschaulichung der vorstehenden Darlegungen wird in Tab. 63 an Hand von konstruierten Zahlen für das normierte Aufkommen und die normierte Bevölkerungszahl von fünf Gemeinden die Berechnung der statistischen Gewichte x und $1 - x$ durchgeführt; die normierten Zahlen werden einmal für wirtschaftlich günstige und sodann für wirtschaftlich ungünstige Verhältnisse konstruiert.

Tab. 63. Berechnung der statistischen Gewichte beim Finanzausgleich.

Nummer der Gemeinde	Gruppe	Normiertes Aufkommen $\bar{a}$	Normierte Bevölkerungszahl $\bar{b}$	$\bar{a}_I - \bar{b}_I$	$\bar{b}_{II} - \bar{a}_{II}$	$(\bar{a}_I - \bar{b}_I)^2 \cdot 10^{-6}$	$(\bar{b}_{II} - \bar{a}_{II})^2 \cdot 10^{-6}$
			Fall a: Wirtschaftlich günstige Verhältnisse				
1	I	580 000	500 000	80 000		6400	
2	I	320 000	300 000	20 000		400	
3	II	70 000	130 000		60 000		3600
4	II	20 000	40 000		20 000		400
5	II	10 000	30 000		20 000		400
Zusammen		1 000 000	1 000 000			6800	4400
			Fall b: Wirtschaftlich ungünstige Verhältnisse				
1	I	510 000	500 000	10 000		100	
2	II	290 000	300 000		10 000		100
3	II	90 000	130 000		40 000		1600
4	I	60 000	40 000	20 000		400	
5	I	50 000	30 000	20 000		400	
Zusammen		1 000 000	1 000 000			900	1700

Fall a:

$$x = \frac{6800}{6800 + 4400} = \frac{6800}{11200} = 0{,}61$$

Fall b:

$$x = \frac{900}{900 + 1700} = \frac{900}{2600} = 0{,}35$$

[1]) Vgl. hierzu F. Burkhardt, Statistische und mathematische Betrachtungen über einige geldliche Ausgleichsprobleme der Verwaltung unter besonderer Berücksichtigung des Finanzausgleichs. Archiv für mathematische Wirtschafts- und Sozialforschung, Bd. I, Heft 2, 1935, S. 65 u. f.

Dritter Abschnitt.

Bezugnahme auf die Wahrscheinlichkeitsrechnung.

103. Die Beziehungen der Statistik zur Wahrscheinlichkeitsrechnung werden in der Ursachenforschung verwendet. In der Statistik sind die Ursachen, die den beobachteten Tatbeständen und Erscheinungen zugrunde liegen, schwer zu durchblicken. Dasselbe ist bei den Ereignissen der Fall, mit welchen sich die Wahrscheinlichkeitstheorie beschäftigt. Frühzeitig schon ist der Blick auf diesen Umstand gerichtet und die Analogie im ersten Anlauf für eine so vollständige gehalten worden, daß man glaubte, den Eintritt männlicher und weiblicher Geburten, das Erleben und Nichterleben eines bestimmten Alters ohne weiteres genau so behandeln zu dürfen wie die Ziehungen aus einer Urne, die mit zwei Arten von Kugeln gefüllt ist. Nur die Art der Füllung hatte zu wechseln je nach den Erfahrungen, die von Fall zu Fall vorlagen. In solcher Weise wandte Laplace im VI. und VIII. Kapitel seiner Théorie analytique des probabilités die Wahrscheinlichkeitsrechnung auf die Geburten und auf die menschliche Lebensdauer, die Dauer von Ehen und anderen Verbindungen an. Erst einer viel späteren Zeit war es vorbehalten, darnach zu fragen, unter welchen Bedingungen ein solcher Vorgang statthaft ist und wie man aus den Erfahrungstatsachen selbst ersehen kann, ob die Bedingungen erfüllt sind. Es hat sich gezeigt, daß die Materien, auf welche die statistischen Methoden angewandt werden, mannigfacher und komplizierter sind als diejenigen, welche sich die Wahrscheinlichkeitsrechnung, zum Teil mit Hypothesen arbeitend, für ihre Entwicklungen zurechtgelegt hat, und daß nur wenige davon ein mit diesen übereinstimmendes oder doch ähnliches Verhalten zeigen. Trotz dieser Einschränkung haben die Ergebnisse der Wahrscheinlichkeitslehre eine noch über die Fälle der Übereinstimmung hinausreichende Bedeutung. Sie gewähren einen wertvollen Anhalt auch dort, wo nicht alle Folgerungen gezogen werden dürfen.

Bei seinem Versuche, die Statistik zu definieren, hat Edgeworth[1]) geradezu die Beziehungen zur Wahrscheinlichkeitsrechnung als dasjenige bezeichnet, worauf sich das wissenschaftliche Interesse der Statistik richtet. Er gibt zu, daß die Analogien, auf die man sich dabei beruft, keine vollständigen sind; es laufe da ein breites hypothetisches Element mit, aber, wie er meint, nicht so breit, wie es bei mancher anerkannten physikalischen Theorie der Fall ist. — In den folgenden Ausführungen werden die grundlegenden Gesichtspunkte dargelegt[2]).

---

[1]) F. Y. Edgeworth, On the application of the Calculus of Probabilities to Statistics Bulletin de l'Institut international de Statistique. Bd. 13, 1909, S. 505.

[2]) Zu dem Inhalte dieses Paragraphs sei hingewiesen auf die gesammelten Abhandlungen von W. Lexis in „Abhandlungen zur Theorie der Bevölkerungs- und Moralstatistik", 1903, auf L. v. Bortkiewicz „Kritische Betrachtungen zur theoretischen Statistik", Conrads Jahrbücher, 3. Folge, Bde. 8, 10, 11 (1894–1896), E. Kamke, Einführung in die Wahrscheinlichkeitstheorie. Leipzig 1932, A. Timpe, Einführung in die Finanz- und Wirtschaftsmathematik. Berlin 1934, S. 147 u. f. und E. Tornier, Wahrscheinlichkeitsrechnung und allgemeine Integrationstheorie. Leipzig und Berlin 1936.

§ 1. Die mittlere quadratische Abweichung.

104. Wenn es sich um die Würfe mit einer „idealen" Münze handelt, die als eine homogene zylindrische Scheibe anzusehen ist, so ist es logisch begründet, bei jedem einzelnen Wurf die Wappenseite mit derselben Aussicht auf Erfüllung zu erwarten wie die andere Seite. Werden also 2, 4, 6, . . . Würfe gemacht, so ist es logisch am besten begründet, in den aus diesen Würfen gebildeten Kollektiven die Verteilung von 1 Wappen und 1 Nicht-Wappen, 2 Wappen und 2 Nicht-Wappen, 3 Wappen und 3 Nicht-Wappen u. s. w. zu erwarten; denn jede andere Verteilung würde auf eine Begünstigung einer der beiden Münzseiten schließen lassen, was der Voraussetzung widerspricht.

Bei einer ungeraden Anzahl von Würfen gibt es keine gleiche Verteilung der beiden Seiten; es steht aber nichts im Wege, gebrochene Häufigkeitszahlen einzuführen und daher bei jeder Wurfzahl n die logisch erschlossene Häufigkeit von Wappen und Nicht-Wappen mit je $\frac{n}{2}$ anzusetzen, wenn man den gebrochenen Häufigkeitszahlen eine entsprechende Deutung gibt.

Dies kann in folgender Weise geschehen. Bei drei Münzwürfen sind folgende Verteilungen von Wappen (W) und von Nicht-Wappen (N) mit den daneben geschriebenen Häufigkeiten von Wappen logisch gleichberechtigt:

WWW	3
WWN	2
WNW	2
NWW	2
WNN	1
NWN	1
NNW	1
NNN	0

hiernach ist die durchschnittliche Häufigkeit von Wappen in einer Wurfserie

$$\frac{1\cdot 3 + 3\cdot 2 + 3\cdot 1 + 1\cdot 0}{8} = \frac{3}{2},$$

entsprechend der halben Wurfzahl. Dasselbe Resultat ergibt sich für Nicht-Wappen.

Man kann zu diesem Resultat auch durch folgende Schlußweise kommen: Macht man zwei Serien von je drei Würfen und vereinigt sie zu einem Kollektiv, so ist in diesem die logisch allein berechtigte Verteilung: 3 Wappen, 3 Nicht-Wappen; folglich entfallen auf eine Serie durchschnittlich die Wiederholungszahlen $\frac{3}{2}$ für Wappen und $\frac{3}{2}$ für Nicht-Wappen.

Die Häufigkeitszahl, die sich für einen Erfolg bei eingliedrigem Kollektiv, also bei einem Wurf, auf logischer Erwägung fußend ergibt, nennt man seine Wahrscheinlichkeit oder seine Chance.

Bei einer Münze ist also die Wahrscheinlichkeit oder Chance für Wappen $p = \frac{1}{2}$, für Nicht-Wappen $q = \frac{1}{2}$; $p + q$ ist in jedem derartigen Falle, wo nur

zwei denkbare Erfolge vorliegen, gleich 1. Sind mehr als zwei Erfolge denkbar, so kann man dem einen von ihnen, auf den man gerade seine Aufmerksamkeit richtet, die Gesamtheit der übrigen als Nichterfolg gegenüberstellen und hat dann wieder $p + q = 1$.

Ist das Objekt der Versuche ein „idealer" Würfel, d. i. ein Körper, der genau die Gestalt eines Würfels hat und dessen Masse homogen oder doch so verteilt ist, daß ihr Schwerpunkt mit dem geometrischen Mittelpunkt zusammenfällt, wird ferner das Fallen einer bestimmten Seite, z. B. der Seite 1, als Erfolg bezeichnet, alles andere als Nichterfolg, so führt die logische Überlegung dazu, dem Erfolg die Wahrscheinlichkeit $\frac{1}{6}$, dem Nichterfolg die Wahrscheinlichkeit $\frac{5}{6}$ zuzuschreiben. Schafft man n Würfe, so ist die logisch erschlossene Verteilung die, daß $n \cdot \frac{1}{6}$ Erfolge und $n \cdot \frac{5}{6}$ Nichterfolge sich einstellen. Bei einem durch 6 teilbaren n sind diese Häufigkeiten auch praktisch möglich, bei jedem andern n bedürfen sie der Deutung. Ist $n = 6$, so ist die Verteilung von Erfolg und Nichterfolg durch 1 und 5 gekennzeichnet. Ist hingegen beispielsweise $n = 4$, so denke man sich 3 solche Serien von je 4 Versuchen ausgeführt und vereinige sie zu einem Kollektiv; in diesem ist die logisch zu erwartende Verteilung 2 Erfolge und 10 Nichterfolge; daher entfallen auf eine Serie durchschnittlich die Häufigkeiten $\frac{2}{3}$ und $\frac{10}{3}$, und das sind in der Tat die Produkte $n \cdot \frac{1}{6}$ und $n \cdot \frac{5}{6}$.

Die andere Art der Deutung ist hier nicht ohne weiteres anwendbar, weil die möglichen Verteilungen

1111	4
1110	3
1101	3
1011	3
0111	3
1100	2
1010	2
1001	2
0110	2
0101	2
0011	2
1000	1
0100	1
0010	1
0001	1
0000	0

nicht mehr logisch gleich berechtigt sind, wie es auch Erfolg (1) und Nichterfolg (0) nicht sind.

105. In Verallgemeinerung der vorgebrachten Beispiele kann folgendes festgestellt werden: Hat ein Erfolg die Wahrscheinlichkeit p, der Nichterfolg die Wahrscheinlichkeit q, so ist in n Versuchen die logisch berechtigte Erwartung die, daß np Erfolge und nq Nichterfolge sich einstellen.

Werden N Serien von je n Versuchen ausgeführt, so sind dementsprechend Nnp Erfolge zu erwarten.

Wie verhalten sich diese logisch begründeten Aussagen zur Wirklichkeit?

Um diese Frage beantworten zu können, sind mit Materien wie die hier betrachteten, also mit Münzen, Würfeln, mit Urnen, die in verschiedener Weise mit gleichen, verschieden bezeichneten oder sonstwie unterschiedenen Kugeln gefüllt waren, u. ä., vielfache Versuche unternommen worden. Das Ergebnis kann in folgenden Sätzen zusammengefaßt werden:

1. Die wirklich eingetretene Häufigkeit des Erfolgs in n Versuchen ist im allgemeinen nicht np, sondern weicht davon mehr oder weniger ab.

2. Bildet man den Durchschnitt der Häufigkeiten in N solchen Serien, so kommt er im allgemeinen dem logisch begründeten Werte np um so näher, je größer die Zahl N ist.

3. Die auf einen Versuch entfallende durchschnittliche Häufigkeit des Erfolgs nähert sich mit wachsender Gesamtzahl der Versuche, d. i. Nn, seiner Wahrscheinlichkeit p.

Die in diesen Aussagen zusammengefaßten Tatsachen der Erfahrung bilden den Inhalt dessen, was man als Gesetz der großen Zahlen bezeichnet.[1]

Als Beleg dazu führen wir einige Versuchsreihen an.

Beispiel 1. Zwölf Würfel wurden 4096mal hingeworfen; als Erfolg wurde es verzeichnet, wenn 4, 5 oder 6 erschien, als Nichterfolg, wenn anders bezeichnete Flächen nach oben lagen. Da Erfolg und Nichterfolg logisch gleichberechtigt sind, so wären bei jedem Wurf 6 Erfolge und bei einem einzelnen Versuch die durchschnittliche Häufigkeit des Erfolgs mit $\frac{1}{2}$ zu erwarten. In Wirklichkeit war das Resultat folgendes:[2]

:folge	in Würfen
0	.
1	7
2	60
3	198
4	430
5	731
6	948
7	847
8	536
9	257
10	71
11	11
12	.
	4096

[1]) P. Flaskämper legt in den Abhandlungen „Die Statistik und das Gesetz der großen Zahlen“, Allg. Stat. Archiv, 16. Band, 1927, S. 501, „Das Problem der Gleichartigkeit in der Statistik“, ebenda 19. Band, 1929, S. 232 und „Die Bedeutung der Zahl für die Sozialwissenschaften“, ebenda 23. Band, 1933, S. 68, dar, daß es Fälle gibt, in denen dieses Gesetz nicht in Betracht zu ziehen ist.

[2]) Encycl. Brit., 10. Aufl., Bd. XXVIII, S. 282.

Bezeichnet man mit X die Anzahl der Erfolge, mit z die Anzahl der Würfe, in welchen sie vorkommen, so hat man es mit der Verteilungstafel eines unstetigen Kollektivs vom Umfang $N = 4096$ zu tun, aus der nach den entwickelten Methoden das arithmetische Mittel von X bestimmt werden kann; es ergibt sich mit

$$M = 6{,}139,$$

welche Zahl mit $12 \cdot \frac{1}{2} = 6$ zu vergleichen ist; ferner die durchschnittliche Häufigkeit in einem Versuch

$$\frac{6{,}139}{12} = 0{,}5116,$$

die mit $\frac{1}{2}$ zu vergleichen ist.

Eine ebenso angeordnete, nur etwas umfangreichere Versuchsreihe ergab folgendes Resultat:[1])

Erfolge	in Würfen
0	1
1	14
2	103
3	302
4	711
5	1231
6	1411
7	1351
8	844
9	391
10	117
11	21
12	3
	6500

Hier steht der Zahl $12 \cdot \frac{1}{2} = 6$ die Zahl $M = 6{,}116$ und der Zahl $\frac{1}{2}$ die Zahl $\frac{6{,}116}{12} = 0{,}5097$ gegenüber. Die Annäherung der aus der Wirklichkeit hergeholten Zahlen an die logisch festgestellten ist eine größere. Das darf jedoch nicht als eine durchgängige Regel betrachtet werden; die Erfahrung weist bloß auf eine „Konvergenz im allgemeinen" hin, bei der es auch wohl vorkommen kann, daß gelegentlich bei Vermehrung der Versuche eine Verminderung der Annäherung sich einstellt.

Übrigens darf nicht übersehen werden, daß die Versuche mit verschiedenen Würfeln gemacht worden sind.

Auf eines ist noch hinzuweisen: Würfe mit 0 und 12 Erfolgen blieben bei der kleineren Reihe aus, fehlen aber nicht bei der längeren.

Beispiel 2. Aus einer Urne mit gleicher Anzahl weißer und schwarzer Kugeln wurden Ziehungen gemacht; die gezogene Kugel wurde nach Feststellung der Farbe zurückgelegt und unter die andern gemengt; als Erfolg gilt eine schwarze

[1]) Mem. and Proc. of the Manchester Lit. and Phil. Soc., Bd. 51, 1907.

Kugel. Es wurde gezählt, wieviel schwarze Kugeln in den aufeinanderfolgenden Gruppen von 2, 3, 4, ... Ziehungen erschienen sind; das Ergebnis war bei den 5- bis 7gliedrigen Gruppen folgendes: [1])

Erfolge	5gliedrige Gruppen	6gliedrige Gruppen	7gliedrige Gruppen
0	30	17	9
1	125	65	34
2	277	166	104
3	224	192	151
4	136	166	148
5	27	69	95
6		8	40
7			4
	819	683	585

Die in den einzelnen Gruppen logisch zu erwartenden Häufigkeiten des Erfolgs sind:

$$5 \cdot \frac{1}{2} = 2{,}5 \qquad 6 \cdot \frac{1}{2} = 3 \qquad 7 \cdot \frac{1}{2} = 3{,}5;$$

sie sind zu vergleichen mit den durchschnittlichen Häufigkeiten aus den Versuchen, nämlich

$$2{,}479 \qquad 2{,}972 \qquad 3{,}470;$$

die daraus durch Division mit 5, 6, 7 abgeleiteten Häufigkeiten für einen Versuch:

$$0{,}4958 \qquad 0{,}4953 \qquad 0{,}4957$$

sind entgegenzuhalten der logisch begründeten Zahl

$$0{,}5.$$

Im vorstehenden ist der Wahrscheinlichkeitsbegriff auf die Häufigkeitsauffassung gegründet worden. Diese Auffassung ist ein Spezialfall der allgemeinen von Tornier[2]) gegebenen mengentheoretischen Begründung des Wahrscheinlichkeitsbegriffs. Tornier geht von den beiden Grundvorstellungen, der Versuchsvorschrift und ihren logisch möglichen Realisierungen aus. Eine Versuchsvorschrift besteht z. B. im Werfen eines Würfels und im Notieren der Ergebnisse. In diesem Beispiel stellt jede Folge, die aus den Ziffern 1, 2, 3, 4, 5 und 6 besteht, eine logisch mögliche Realisierung dar. Die Gesamtheit aller logisch möglichen Realisierungen bezeichnet Tornier als Folgenmenge. Die Gesamtheit aller Realisierungen setzt sich aus Teilgesamtheiten (Grundmengen) von Realisierungen zusammen, von denen jede durch eine Eigenschaft charakterisiert ist. So bilden in unserem Beispiel alle Realisierungen, die mit 1, 3, 5 beginnen, eine Grundmenge. Den Grundmengen werden Zahlenwerte (Wahrscheinlichkeiten) zugeordnet (z. B. die

[1]) A. Quetelet, Lettres sur la théorie des probabilités, Bruxelles 1846, S. 374.

[2]) E. Tornier, Wahrscheinlichkeitsrechnung und allgemeine Integrationstheorie. Leipzig und Berlin 1936, S. 101 u. f.

Wahrscheinlichkeit dafür, daß eine Realisierung mit 1, 3, 5 beginnt). Das System der so bewerteten Mengen einschließlich der Bewertungen bezeichnet Tornier als das durch die Versuchsvorschrift induzierte Wahrscheinlichkeitsfeld. Die Wahrscheinlichkeiten selbst können entweder apriorisch erschlossen werden, oder sie sind auf induktiv-statistischem Wege zu bestimmen. In der mathematischen Wahrscheinlichkeitstheorie gilt es, in diesen Wahrscheinlichkeitsfeldern die Verknüpfungsregeln in Bezug auf das System der bewerteten Mengen und in Bezug auf die Bewertungen aufzustellen. Man erkennt ohne weiteres, daß diese allgemeine mengentheoretische Auffassung die Häufigkeitsauffassung als Spezialfall in sich schließt.

106. Die logischen Erwägungen lassen sich weiter führen in folgender Richtung. Es ist gesagt worden, daß die Häufigkeiten des Erfolgs in den N Serien zu je n Versuchen von der logisch begründeten Erwartung np in der Regel abweichen; es läßt sich nun auch die zu erwartende mittlere Abweichung logisch feststellen und kann dann mit der aus der Erfahrung abgeleiteten verglichen werden.

Für $n=1$ wird man zu folgender Überlegung geführt: Der Erfolg kann die Häufigkeit 1 oder 0 haben (d. h. er kann eintreten oder nicht); die Abweichungen der logisch begründeten Häufigkeit p beträgt $1-p$, bzw. $-p$; ersteres ist mit der Häufigkeit p, letzteres mit der Häufigkeit q zu erwarten, infolgedessen ist das Quadrat der zu erwartenden mittleren Abweichung

$$p\,(1-p)^2+qp^2=p\,q^2+qp^2=p\,q.$$

Bei n Versuchen ($n \geqq 1$), die sich als eine Summe von n Einzelversuchen darstellen, tritt bei Unabhängigkeit der Variablen eine Summierung der Quadrate der mittleren Abweichungen ein (Art. 83), was hier gleichbedeutend ist mit der Vervielfachung des für den Einzelversuch gefundenen Resultates. Also ist die logisch erschlossene mittlere Abweichung in einer Reihe von n Versuchen durch die Formel

$$\mu_{(n)}=\sqrt{n\,p\,q} \qquad (1)$$

bestimmt.

Aus der Häufigkeit des Erfolgs in einer solchen Reihe wird die Häufigkeit des Erfolgs in einem Versuch durch Division mit n gefunden; durch denselben Prozeß wird aus $\mu_{(n)}$ die auf einen Versuch bezügliche mittlere Abweichung gewonnen:

$$\mu_{(1)}=\sqrt{\frac{p\,q}{n}}. \qquad (2)$$

Die Größe $\mu_{(n)}$ wächst mit der Versuchszahl im Verhältnis ihrer Quadratwurzel, die Größe $\mu_{(1)}$ nimmt in demselben Verhältnis ab.

Die Erprobung dieser Formeln an dem im vorigen Artikel vorgeführten Versuchsmaterial ergibt die folgenden Resultate:

Die erste Versuchsreihe in Beispiel 1 (Art. 105) liefert

$$\mu_{(12)}=1{,}712,$$

während formelgemäß

$$\mu_{(12)}=\sqrt{12\cdot\frac{1}{2}\cdot\frac{1}{2}}=\sqrt{3}=1{,}732$$

gefunden wird. Die zweite dort verzeichnete Versuchsreihe liefert gegenüber dem gleichen theoretischen Werte, nämlich 1,732, den Wirklichkeitswert 1,733.

Für die drei in Beispiel 2 (Art. 105) angeführten Versuchsreihen findet man

$$\mu_{(5)} = 1{,}141 \qquad \mu_{(6)} = 1{,}265 \qquad \mu_{(7)} = 1{,}399,$$

während a priori die Werte

$$\frac{\sqrt{5}}{2} = 1{,}118 \qquad \frac{\sqrt{6}}{2} = 1{,}225 \qquad \frac{\sqrt{7}}{2} = 1{,}323$$

zu erwarten wären.

Man kann also in allen Fällen von einer guten Bestätigung der logisch begründeten Aussagen durch die Wirklichkeit sprechen.

107. Die Bedeutung der vorstehenden Betrachtungen über Zufallsreihen für die praktische Anwendung wird sich aus folgenden Erwägungen ergeben.

Die mit zunehmender Zahl der Versuche sich steigernde Annäherung der Wirklichkeitsergebnisse an die auf logischer Grundlage beruhenden Voraussagen kann nur so gedeutet werden, daß die festen, bleibenden Unterlagen, wie die Eigenschaften der Münzen, der Würfel, der Kugeln in der Urne, gegenüber den unübersehbaren, beständig wechselnden sonstigen Vorgängen, die sich bei den Versuchen, also bei dem Aufwerfen der Münze oder des Würfels, bei den Ziehungen aus der Urne und den Lagenwechseln in derselben abspielen, immer mehr die Oberhand gewinnen. Mit wachsender Zahl der Versuche nähern sich die Ergebnisse in den Grundzügen denjenigen Ergebnissen, die sich einstellen würden, wenn alles streng nach den logischen Voraussagen vor sich ginge. Jene an sich wenig belangreichen, zahlreichen und wechselnden Umstände treten also, je länger man die Versuche fortsetzt, in ihrer Gesamtwirkung immer mehr zurück; sie sind es aber auch, welche das Bild, das nach den logischen Gründen zu erwarten wäre, trüben, entstellen; sie bilden so gewissermaßen die Ursache für die Abweichungen zwischen „Theorie" und „Wirklichkeit".

Wenn wir nun an eine Materie herantreten, die sich der Forschung darbietet, sei es eine anthropologische, biologische, ökonomische oder dergleichen, so wissen wir über die Ursachen, die sie beherrschen, entweder gar nichts oder doch viel zu wenig, um irgend welche Aussagen a priori über ihre Struktur, über zu erwartende Verteilungen machen zu können. Aber die Möglichkeit ist nicht ausgeschlossen, daß die Materie ebenso wie die Zufallsversuche von so durchschlagenden bleibenden Umständen beherrscht wird, daß in einem umfangreichen, aus der betreffenden Materie gebildeten Kollektiv eben diese Umstände wie dort zum überragenden Ausdruck kommen gegenüber den zahlreichen anderen und variierenden Umständen, mit einem Wort: daß sich die Materie ähnlich verhält wie Zufallsergebnisse.

Nach neueren Untersuchungen von van der Waerden[1]) ist in den Fällen, in denen über die Häufigkeit des Erfolgs keine apriorische Aussage gemacht werden kann, nach folgenden Formeln zu rechnen:

[1]) B. L. van der Waerden, „Messung von Wahrscheinlichkeiten, insbesondere Mortalität von Krankheiten, Operationen usw." Berichte der mathematisch-physikalischen Klasse der Sächsischen Akademie der Wissenschaften zu Leipzig. LXXXVIII. Bd., Leipzig 1936, S. 21 u. f. und „Über die richtige Auswertung von Erfolgsstatistiken". Klinische Wochenschrift. Jahrg. 15, 1936, Nr. 47, S. 1718 u. f.

$$p = \frac{m+1}{n+2} \tag{3}$$

$$\mu_{(n)} = n\sqrt{\frac{p(1-p)}{n+3}} \tag{4}$$

$$\mu_{(1)} = \sqrt{\frac{p(1-p)}{n+3}} \tag{5}$$

Hierbei bedeuten n die Anzahl der Versuche in einer Reihe, m die Anzahl der empirisch bestimmten Erfolge bei diesen n Versuchen, p den Mittelwert der unbekannten Wahrscheinlichkeit η für das Eintreten des Erfolges, $\mu_{(n)}$ bzw. $\mu_{(1)}$ die mittlere quadratische Abweichung der absoluten bzw. relativen Häufigkeitszahlen für das Eintreten des Erfolges. Die mathematische Herleitung der Formeln (3), (4) und (5) wird in Art. 129 mit Bezugnahme auf die Untersuchungen von van der Waerden gegeben werden. In der praktischen statistischen Forschung ist darauf zu achten, daß die Zahl der Beobachtungen n unter denselben Voraussetzungen genügend groß ist. Ob dies der Fall ist, kann vielfach durch sachliche Überlegungen beurteilt werden.

Stimmt dann das nach der Formel (4) oder (5) abgeleitete μ mit demjenigen genügend überein, das sich nach den vorgetragenen Methoden aus der beobachteten Verteilung ergibt, so spricht dies für die Zufallsnatur der Materie; eine apodiktische Aussage kann man auch bei noch so guter Übereinstimmung nach einmaliger Prüfung niemals machen.[1])

Liegen über einen und denselben Erfolg zwei Versuchsreihen vor und führen sie zu verschiedenen empirischen Bestimmungen von p, wie das ja fast ausnahmslos der Fall sein wird, so braucht diese Verschiedenheit nicht in den Grundlagen ihre Ursache zu haben; sie kann auch von den zufälligen Schwankungen herrühren, die in den beiden Reihen in ungleichem Maße zur Geltung gekommen sind. Auch dafür, ob das eine oder das andere eher anzunehmen ist, gibt es einen Anhalt. Sind p_1, p_2 die beiden empirischen Bestimmungen für das vermeintlich bestehende p, μ_1, μ_2 ihre ebenfalls empirisch (aus der Verteilung) bestimmten mittleren Abweichungen, so hat die Differenz $|p_1 - p_2|$ die mittlere Abweichung (Art. 83, b))

$$\mu = \sqrt{\mu_1^2 + \mu_2^2};$$

ist $|p_1 - p_2| < 3\mu$, so kann die Differenz eine bloße Folge zufälliger Schwankungen sein; ist dagegen $|p_1 - p_2| > 3\mu$, dann ist mit mehr Berechtigung anzunehmen, daß die Grundlagen in den Materien voneinander abweichen. Die mathematische Begründung hierfür wird in Art. 124 gegeben werden.

[1]) Auf dem Gedanken der Vergleichung des nach der Zufallstheorie a priori bestimmten μ mit jenem, das sich aus der tatsächlich beobachteten Verteilung ergibt, beruht die Lexissche Lehre von der Stabilität oder Dispersion statistischer Reihen. Der Quotient aus dem zweiten Wert durch den ersten ist unter dem Namen Divergenzkoeffizient (Dormoy) oder Fehlerrelation (v. Bortkiewicz) als Maß der Dispersion eingeführt worden. Näheres hierüber findet man bei W. Lexis, Abhandlungen zur Theorie der Bevölkerungs- und Moralstatistik, Jena 1903, S. 170 u. f., wo auch die älteren Lexisschen Arbeiten zu diesem Gegenstande angegeben sind. Neuerdings hat E. Kamke die Lexissche Dispersionstheorie eingehend wahrscheinlichkeitstheoretisch behandelt und die Ergebnisse seiner Forschungen in dem Buch „Einführung in die Wahrscheinlichkeitstheorie" (Leipzig 1932, S. 161 u. f.) niedergelegt.

108. Bevor man jedoch an eine solche Untersuchung schreitet, empfiehlt es sich zu überlegen, ob bei der zu untersuchenden Materie die Voraussetzungen, die den logischen Erwägungen zugrunde liegen, als in genügendem Maße erfüllt anzusehen sind.

Denken wir beispielsweise an die Versuche mit den 12 Würfeln, so wird man zu einer Anwendung der apriorischen Formeln nur dann schreiten, wenn alle Versuche mit demselben Würfelsatz oder doch mit gleichen Würfelsätzen gemacht sind. Der apriorische Wert von p setzt voraus, daß jeder Würfel ein „vollkommener" Würfel sei in dem früher erklärten Sinne. Kann letzteres nicht mit Sicherheit behauptet werden, dann ist es besser, für p den aus den Versuchen abgeleiteten empirischen Wert statt des apriorischen bei der Ausrechnung der Formel zu gebrauchen. Wären aber die Würfe mit verschiedenen Würfelsätzen gemacht, innerhalb deren sich auch unvollkommene Würfel verschiedenen Unvollkommenheitsgrades befinden, dann wäre für die Anwendung der Formeln die Grundlage nicht vorhanden.

Die Nutzanwendung hiervon ist die folgende. Bestimmt man beispielsweise aus mehreren gleich umfangreichen Gesamtheiten von Lebenden das Sterblichkeitsverhältnis, so hat die Anwendung der Formeln auf die Resultate nur dann einen Sinn, wenn die Gesamtheiten in jenen Belangen gleichartig sind, die auf die Sterblichkeit Einfluß haben. Unbegründet wäre sie jedenfalls, wenn die Gesamtheiten aus Personen verschiedenen Geschlechtes, verschiedenen Alters ungleichartig zusammengesetzt sind, wenn neben gesunden auch kranke Personen darin vorkommen.

Ein anderer Fall ist der folgende. Gegenstand der Untersuchung sei eine als Wahrscheinlichkeit auffaßbare Verhältniszahl, die sich auf die Ernteergebnisse der einzelnen Parzellen eines Versuchsfeldes bezieht. Standen die Pflanzen unter ungleichen Wachstumsbedingungen oder war der Boden ungleichwertig oder traf beides zusammen, so wird die nach der theoretischen Formel berechnete mittlere Abweichung der Ernteergebnisse der einzelnen Parzellen im allgemeinen nicht übereinstimmen können mit der aus der Verteilung ermittelten.

Die logische Überlegung geht des weiteren so vor, als ob zwischen den Versuchen und den Objekten, mit welchen sie ausgeführt werden, völlige Unabhängigkeit bestünde. Das kann bei Würfen mit einem Würfel angenommen werden, da der Erfolg eines Wurfes keinen Einfluß auf den Erfolg des nächsten hat, sowie auch er nicht beeinflußt war von dem vorangehenden; darin liegt die Unabhängigkeit. Sie ist hingegen nicht vorhanden beispielsweise bei Ziehungen aus einer Urne, die ursprünglich mit weißen und schwarzen Kugeln in einem bestimmten Zahlenverhältnis gefüllt war; denn das Farbenverhältnis ändert sich im Laufe der Ziehungen, wenn die gezogenen Kugeln nicht wieder hinein gelegt werden, und es hängt somit die Chance einer schwarzen Kugel vor einem bestimmten Zuge davon ab, was die vorangehenden Züge ergeben haben. Die Unabhängigkeit kann hier dadurch hergestellt werden, daß man die gezogene Kugel jedesmal wieder zurücklegt und unter die andern mengt; angenähert wenigstens besteht sie, wenn die Gesamtzahl der Kugeln in der Urne sehr groß ist im Vergleich zur Zahl der Ziehungen. Man müßte also, um Ziehungsserien nach den Formeln zu behandeln, wissen, daß die Zusammensetzung der Urne durch die ganze Dauer der Ziehungen dieselbe blieb (oder sich nur äußerst wenig, wie in dem letztgedachten Falle, geändert hat).

In den praktischen Anwendungen bildet die Frage, ob die Voraussetzung der Unabhängigkeit erfüllt ist, einen der schwierigsten Punkte. Man wird in einem gegebenen Falle eine vorhandene Abhängigkeit wohl feststellen, wird aber kaum je mit Sicherheit behaupten können, daß keine wie immer geartete Abhängigkeit bestehe. Bei den zahllosen, zum Teil noch unerkannten Fäden, die zwischen den Dingen der uns umgebenden Welt und dem darin sich abspielenden Geschehen hin und her gehen, wird es völlige Unabhängigkeit nur selten geben. Bis zu einem gewissen Grade ist es also Hypothese, wenn man in diesem oder jenem Falle von Unabhängigkeit spricht.

109. Wenn es unter den Materien, mit welchen sich die praktische Statistik beschäftigt, kaum eine gibt, bei der die im vorigen aufgeführten Bedingungen in aller Strenge erfüllt sind, so kennt man doch auch manche, die ein Verhalten zeigen, das demjenigen von Zufallsereignissen sehr nahe kommt. Namentlich finden sich solche Materien auf biologischem Gebiete, wo es sich um „rein natürliche“ Vorgänge handelt, um Vorgänge, auf die menschlicher Wille keinen entscheidenden Einfluß nehmen kann. Wirtschaftliche Gegenstände, die solchem Einfluß meist in hohem Maße unterliegen, zeigen selten ein dem Zufälligen nahekommendes Gepräge.

Die folgenden zwei Beispiele gehören in die erste Gruppe.

a) Das Geschlechtsverhältnis der ehelich Lebendgeborenen. Es ist schon durch vielfache Erfahrungen bestätigt, daß sich die männlichen und weiblichen Geburten in langen Geburtenreihen sehr nahe so verhalten wie die Ziehungsergebnisse aus einer Urne mit bestimmter gleichbleibender Füllung mit weißen und schwarzen Kugeln. Wenn also in einer Anzahl gleich umfangreicher Geburtenfolgen das Verhältnis der männlichen Geburten zur Gesamtzahl bestimmt wird, so wird die mittlere Abweichung dieser Reihe von Verhältniszahlen, mit $\sqrt{\frac{pq}{n+3}}$ verglichen, eine um so größere Übereinstimmung zeigen, je größer der Umfang n einer solchen Folge ist; der Wert p des erwähnten Verhältnisses läßt sich logisch nicht erschließen, man nimmt dafür den aus der Beobachtung abgeleiteten Wert p, der nach Formel (3) zu berechnen ist. In Formel (3) ist m die Zahl der ehelich lebendgeborenen Knaben und n die Gesamtzahl der ehelich Lebendgeborenen.

So einfach, wie hier angenommen, liegen die Umstände in der Wirklichkeit nicht, die Rechnung muß ihnen erst angepaßt werden.

Wir wollen die ehelich Lebendgeborenen des Jahres 1931 in den 29 deutschen Ländern und Landesteilen [1]) einer Untersuchung auf das Geschlechtsverhältnis unterziehen und bestimmen für jedes Gebiet die Zahl der Knaben unter 1000 ehelich Lebendgeborenen (Knabenquote). Die betreffenden Zahlen bewegen sich zwischen den Grenzen 490,7 und 534,5; sie stützen sich auf verschiedene Geburtenmengen, die zwischen 662 und 116 099 schwanken.

Der eine Rechnungsweg bestünde darin, daß man die Gebiete nach ihrer Geburtenmenge in Klassen einteilt, zu jeder Klasse die Verhältniszahl bestimmt und diese Verhältniszahlen als Kollektivreihe behandelt. Dabei wird im Sinne der Theorie jede einzelne Verhältniszahl der mittleren Geburtenmenge der betreffenden Klasse zugeordnet, was um so eher statthaft ist, je mehr Gebiete in eine Klasse fallen und je gleichmäßiger die Geburtenmengen über die Klasse verteilt sind.

[1]) Statistik des Deutschen Reichs, Bd. 441, 1934, S. 24.

Die Anwendung dieses Verfahrens schließt das vorliegende Material aus, weil die Anzahl der Gebiete zu klein und ihre Geburtenmengen zu ungleichmäßig verteilt sind: Bei 1 Gebiet liegt die Geburtenzahl unter 1000, bei 7 Gebieten zwischen 1000 und 10 000, bei 19 zwischen 10 000 und 100 000 und bei 2 über 100 000.

Eine andere Klasseneinteilung wäre die nach den Werten der Verhältniszahl; diese geht von 490,7 bis 534,5, und ihre Werte verteilen sich auf die Gebiete, wie aus der nebenstehenden Tabelle ersichtlich ist.

Tab. 64. Knabenquote der ehelich Lebendgeborenen.

Knabenquote	Zahl der Gebiete
490—495	1
495—500	1
500—505	1
505—510	1
510—515	11
515—520	11
520—525	2
525—530	—
530—535	1
	29

Auch diese Verteilung ist recht unregelmäßig und hat den Nachteil, daß in einer Klasse Gebiete sehr ungleicher Größe vereinigt sind. Man findet daraus

$$M = 514{,}052,$$

was 105,78 Knabengeburten auf 100 Mädchengeburten entspricht, und

$$\mu = 7{,}323.$$

Faßt man alle Gebiete zusammen, so ergeben sich 910 545 ehelich Lebendgeborene, darunter 469 395 Knaben, was zu einem Verhältnis von

$$515{,}510$$

Knaben unter 1000 ehelich Lebendgeborenen (106,40 Knaben auf 100 Mädchen) führt; rechnet man auf Grund von (3) also mit $p = 0{,}51551$, $q = 0{,}48449$ nach der theoretischen Formel (4), so erhält man für $\mu_{(n)}$ den Wert

$$\mu_{(n)} = 1000\sqrt{\frac{0{,}51551 \cdot 0{,}48449}{1000 + 3}} = 15{,}780,$$

der erheblich größer ist als der aus der Verteilungstafel abgeleitete. Bei der Berechnung von $\mu_{(n)}$ auf Grund der Formel (4) wird $n = 1000$ gesetzt, weil die Bestimmung von μ nach der Verteilungstafel auf eine Geburtenzahl von 1000 eingestellt ist.

In einem Falle wie der vorliegende ist es jedoch besser, die theoretische Berechnung von μ auf eine andere Grundlage zu stellen. Bezeichnet man mit n_1, $n_2, \ldots n_{29}$ die um 3 vermehrte Zahl der ehelich Lebendgeborenen in den einzelnen Gebieten und behält man für p und q die obige Bedeutung bei, so sind die mittleren Abweichungsquadrate der einzelnen Gebiete

$$\frac{pq}{n_1}, \frac{pq}{n_2}, \ldots \frac{pq}{n_{29}}$$

mithin ist das mittlere Quadrat der Abweichung für ein Gebiet auf Grund aller $N = 29$ Gebiete

$$\frac{1}{N}\left(\frac{pq}{n_1} + \frac{pq}{n_2} + \cdots + \frac{pq}{n_{29}}\right)$$

$$= pq\,\frac{1}{N}\left(\frac{1}{n_1} + \frac{1}{n_2} + \cdots + \frac{1}{n_{29}}\right) = \frac{pq}{H}$$

wenn man mit H das harmonische Mittel der um 3 vermehrten Geburtenzahlen in den einzelnen Gebieten bezeichnet.

Hiernach ist die theoretische mittlere Abweichung der Verhältniszahl (auf 1000 Geburten gerechnet)

$$\mu_0 = 1000 \sqrt{\frac{pq}{H}}.$$

Nach dieser Formel ergibt sich, da sich H mit 5880 berechnet,

$$\mu_0 = 6{,}52$$

in weit besserer Übereinstimmung mit dem aus der Verteilungstafel berechneten $\mu = 7{,}323$.

b) Das Geschlechtsverhältnis der im ersten Lebensjahre ehelich Gestorbenen. Die Statistik für dasselbe Jahr und für die gleichen Gebiete, auf die sich die Geburtenzahlen des vorigen Beispiels beziehen, gibt auch Aufschluß über die Sterblichkeit der beiden Geschlechter im ersten Lebensjahre[1]). Wird diese Statistik in derselben Weise verwertet wie die vorige, d. h. werden die Verhältniszahlen der im ersten Lebensjahr ehelich gestorbenen Knaben unter 1000 ehelich Gestorbenen (Knabenquote) in Klassen eingeteilt, so ergibt sich nachstehende Verteilungstafel. Aus ihr berechnen sich das arithmetische Mittel

$$M = 570{,}172$$

und die mittlere Abweichung

$$\mu = 27{,}870.$$

Tab. 65. Knabenquote der im ersten Lebensjahr ehelich Gestorbenen.

Knabenquote	Zahl der Gebiete
480—490	1
490—500	—
500—510	1
510—520	—
520—530	—
530—540	—
540—550	1
550—560	6
560—570	4
570—580	5
580—590	7
590—600	1
600—610	2
610—620	—
620—630	—
630—640	1
	29

Die Streuung ist hier wesentlich größer als bei den Geburten, wie es schon die Tafel zeigt; das Verhältnis stellt sich für das männliche Geschlecht erheblich ungünstiger als bei den Geburten, indem 132,65 Knaben auf 100 Mädchen entfallen; es ist dies der ziffernmäßige Ausdruck für die bekannte Tatsache, daß das Übergewicht des männlichen Geschlechtes bei der Geburt durch seine stärkere Sterblichkeit im Lebensbeginn wieder ausgeglichen wird.

Aus der Zusammenfassung aller Gebiete: 70 668 Gestorbene, davon 40 703 männlich, ergibt sich die Verhältniszahl

$$575{,}97,$$

was in anderer Weise ausgedrückt 135,8 Knaben auf 100 Mädchen entspricht.

[1]) Statistik des Deutschen Reichs, Bd. 441, 1934, S. 51.

Mit $p = 0{,}57597$, $q = 0{,}42403$ ergibt sich der theoretische Wert der mittleren Abweichung

$$\mu_{(n)} = 1000 \sqrt{\frac{0{,}57597 \cdot 0{,}42403}{1000 + 3}} = 15{,}60,$$

also bedeutend kleiner als aus der obigen Verteilung. Das hat wieder seinen Grund in der großen Verschiedenheit der Grundzahlen, aus welchen die einzelnen Verhältniszahlen abgeleitet sind, also in der großen Ungleichheit der Gebiete.

Bringt man diese in Rechnung, indem man sich des harmonischen Mittels der Zahlen der Gestorbenen

$$H = 410{,}6$$

bedient, so ergibt sich eine mittlere Abweichung

$$\mu_0 = 1000 \sqrt{\frac{pq}{H}} = 24{,}84,$$

die mit dem aus der Verteilungstafel abgeleiteten Werte 27,87 gut übereinstimmt.

Die Untersuchung zeigt, daß die Verteilung der Geschlechter auf die Geburten einem zufälligen Geschehen besser entspricht, als ihre Verteilung auf die Sterbefälle des ersten Lebensjahres.

110. Die Formeln $\sqrt{npq}$ und $\sqrt{\frac{pq}{n}}$, erstere auf die absolute, letztere auf die relative Häufigkeit eines Vorkommens in n Versuchen oder Beobachtungen bezüglich, finden vielfältige Anwendung bei verschiedenen wissenschaftlichen Untersuchungen, wenn die Werte p und q auf Grund von apriorischen Erwägungen gegeben sind. Lassen sich die Werte p und q nicht apriorisch erschließen, sondern werden sie auf Grund von Beobachtungen nach Formel (3) berechnet, dann kommen für die Bestimmung der mittleren quadratischen Abweichung die Formeln $n\sqrt{\frac{pq}{n+3}}$ und $\sqrt{\frac{pq}{n+3}}$ in Betracht. Zwei Fragestellungen sind besonders hervorzuheben.

I. Über eine Materie ist auf Grund apriorischer Erwägungen eine Hypothese aufgestellt worden in dem Sinne, daß eine Aussage über die Häufigkeit des Vorkommens von A in einer Anzahl von Fällen gemacht ist. Ist eine über die betreffende Materie angestellte Beobachtungsreihe als eine Bestätigung der Hypothese anzusehen?

Angenommen p sei die a priori aufgestellte relative Häufigkeit von A, dann wäre A in n Beobachtungen npmal zu erwarten; in Wirklichkeit sei es $a$mal aufgetreten.

Wenn die Verteilung von A und seines Gegensatzes α in den n beobachteten Fällen außer von den grundlegenden Umständen vom Zufall beeinflußt wäre, so wäre aus diesem letzteren Einfluß eine mittlere Abweichung des a von np im Betrage $\sqrt{npq}$ zu erwarten, oder auch eine mittlere Abweichung der relativen Häufigkeit $\frac{a}{n}$ von p im Betrage $\sqrt{\frac{pq}{n}}$. Macht die tatsächlich eingetretene

Abweichung $|a - np|$, bzw. $\left|\frac{a}{n} - p\right|$, das Dreifache[1]) oder ein noch größeres Vielfaches von $\sqrt{npq}$, bzw. $\sqrt{\frac{pq}{n}}$ aus, dann ist nicht anzunehmen, daß dies die Wirkung des Zufalls allein sein könnte, vielmehr ist dann eher auf ein Nichtzutreffen der zugrundegelegten Hypothese zu schließen.

Ein endgültiges Urteil wird man jedoch auf eine einzelne, wenn auch noch so umfangreiche Beobachtungsreihe nicht stützen wollen. Erst wenn mehrere vorliegen, wenn die Abweichungen wechselnde Richtung und ein Ausmaß aufweisen, welches über das Dreifache von $\sqrt{npq}$, bzw. $\sqrt{\frac{pq}{n}}$ nicht hinausgeht, wird man mit Berechtigung von einem Zutreffen der Hypothese sprechen können. Daß dabei zweifelhafte Fälle auftreten können, ist selbstverständlich; nur eine Erweiterung der Erfahrungsgrundlage kann dann Aufklärung verschaffen.

Erstes Beispiel. Bei 20 000 Versuchen, wobei mit zwei durch Färbung unterschiedenen Würfeln geworfen wurde, erhielt R. Wolf[2]) folgende Wiederholungszahlen der sechs Würfelseiten:

	1	2	3	4	5	6
Weißer Würfel.....	3246	3449	2897	2841	3635	3932
Roter Würfel......	3407	3631	3176	2916	3448	3422

Vertragen sich diese Ergebnisse mit der Hypothese gleichmöglicher Fälle, nach welcher für jede Seite die logisch erschlossene Wahrscheinlichkeit oder relative Häufigkeit $\frac{1}{6}$ wäre?

Dieser Hypothese entspräche die durchgängige Wiederholungszahl 3333 $\left(= \frac{1}{6} \cdot 20\,000\right)$; die Abweichungen davon betragen beim

weißen Würfel.....	-87	116	−436	−492	302	599
roten Würfel......	74	298	−157	−417	115	89;

die mittlere Abweichung ist $\sqrt{20\,000 \cdot \frac{1}{6} \cdot \frac{5}{6}} = 52{,}7$, ihr Dreifaches 158,1. Beim weißen Würfel geht die größte Abweichung über das 11fache hinaus, beim roten beträgt sie fast das 8fache; so große Abweichungen sind aber als zufällig nicht zu erwarten; die Versuche sprechen also dafür, daß bei den Würfeln die Annahme gleicher Möglichkeit der Fälle, also die Voraussetzung ihrer „Vollkommenheit", weitaus nicht zutrifft.

Zweites Beispiel. Durch Kreuzung gelbkörniger mit grünkörnigen Erbsenrassen gewann Mendel neue Pflanzen, von denen er 8023 Samen erntete; davon waren 6022 gelb und 2001 grün. Dieses Ergebnis kann als eine Bestätigung der

[1]) Die nähere Darlegung dieser Regel findet sich in Art. 124.

[2]) Dritte Mitteilung über eine neue Reihe von Würfelversuchen. Vierteljahrsschr. der Naturforschenden Gesellschaft in Zürich, 27. Jahrg., 1882, S. 242 u. 243.

Mendelschen Lehre betrachtet werden, nach der $\frac{3}{4}$ gelbe und $\frac{1}{4}$ grüne Samen zu erwarten sind, wenn das Merkmal gelb dominant und grün rezessiv ist[1]).

Die Abweichungen der beobachteten von der hypothetischen Verteilung betragen, da $\frac{3}{4} \cdot 8023 = 6017{,}25$ und $\frac{1}{4} \cdot 8023 = 2005{,}75$; 4,75 und $-4{,}75$. Mit

$$\sqrt{8023 \cdot \frac{3}{4} \cdot \frac{1}{4}} = 39$$

verglichen, ergibt sich, daß die Abweichungen zufälliger Natur sind und daß die Beobachtung mit der Hypothese gut vereinbar ist.

Wiederholte, von verschiedenen Forschern an verschiedenen Orten ausgeführte Versuche über dieselbe Materie haben das Zutreffen der Mendelschen Lehre in dem vorliegenden Fall so gut wie außer Zweifel gestellt. Eine große Versuchsreihe hat Bateson veröffentlicht; sie umfaßt 15806 geerntete Samen, die sich gemäß den Zahlen 11903 und 3903 auf die Farben gelb und grün verteilten, so daß die Anteile der beiden Farben sich mit 0,7531 und 0,2469 ergeben[2]). Dem steht die nach der Zufallstheorie zu erwartende mittlere Abweichung

$$\sqrt{\frac{\frac{3}{4} \cdot \frac{1}{4}}{15806}} = 0{,}0034$$

gegenüber, so daß die beobachtete weitaus innerhalb jener Grenzen liegt, die noch als zulässig gelten können.

Drittes Beispiel. Durch Kreuzung einer Zwergrasse mit einer hochwachsenden Erbsenrasse erhielt Mendel 1064 Nachkommen, von welchen 787 dem hohen, 277 dem Zwergtypus angehörten[3]).

Nach derselben angenommenen Gesetzmäßigkeit waren 798 Pflanzen der ersten und 266 Pflanzen der zweiten Art zu erwarten; die Abweichungen der empirischen Werte von den theoretischen sind -11 und $+11$. An

$$\sqrt{1064 \cdot \frac{3}{4} \cdot \frac{1}{4}} = 14{,}1$$

gemessen, erweisen sie sich als klein genug, um als zufällige Störungen gelten zu können.

Die mathematischen Grundlagen der Mendelschen Lehre sind eingehend von P. Riebesell (Biometrik nnd Variationsstatistik, Handbuch der biologischen Arbeitsmethoden, Abteilung V, Teil 2, 1. Hälfte, 1924, S. 817 u. f.), E. Weber (Variations- und Erblichkeitsstatistik, München 1935, S. 188 u. f.), F. Ringleb (Mathematische Methoden der Biologie, Leipzig und Berlin 1937, S. 73 u. f.. S. 145 u. f.) behandelt

[1]) Vgl. P. Riebesell, Mathematische Statistik und Biometrik. Frankfurt a. M. und Berlin 1932, S. 51 u. f.

[2]) W. Johannsen, Elemente der exakten Erblichkeitslehre, 3. Aufl., Jena 1926. S. 462.

[3]) W. Johannsen, Elemente der exakten Erblichkeitslehre, 3. Aufl., Jena 1926, S. 427.

worden. Die binominale Verteilung, deren Kenntnis hierbei vorausgesetzt wird, wird in § 2 behandelt werden. Vgl. hierzu auch H. Münzner, Über die Schnelligkeit der Rassenmischung. Archiv für mathematische Wirtschafts- und Sozialforschung. Bd. I, Heft 1, 1935, S. 36 u. f.

II. Eine in gewissem Sinne zu der vorigen entgegengesetzte Fragestellung ist es, wenn man die Aufgabe so wendet: Aus zwei Beobachtungsreihen der Umfänge n_1, n_2 haben sich nach (3) auf Grund des Vorkommens von A zwei verschiedene Werte p_1, p_2 ergeben. Wird diese Verschiedenheit als eine wirkliche angenommen, darf man dann erwarten, daß sie bei einer zweiten Erhebung von derselben Beschaffenheit verschwindet, d. h. von zufälligen Störungen aufgehoben wird?

Nimmt man p_1, p_2 für die nach (3) berechneten Mittelwerte in den beiden Reihen, so sind die diesen Reihen theoretisch zukommenden mittleren Abweichungen $\sqrt{\frac{p_1 q_1}{n_1+3}}$, $\sqrt{\frac{p_2 q_2}{n_2+3}}$; daraus ergibt sich nunmehr als mittlere Abweichung des Unterschiedes $|p_1 - p_2|$

$$\mu = \sqrt{\frac{p_1 q_1}{n_1+3} + \frac{p_2 q_2}{n_2+3}}; \tag{6}$$

beträgt $|p_1 - p_2|$ das Drei- oder Mehrfache davon, so ist nicht zu erwarten, daß durch zufällige Abweichungen in einem neuen Falle derselben Art der Unterschied verwischt wird, daß sich also aus beiden Reihen gleiche Werte für p $(p_1 = p_2)$ ergeben.

Erstes Beispiel. Nach der amtlichen deutschen Statistik für das Jahr 1933 [1]) betrug die Zahl der

	ehelichen	unehelichen
Geborenen	$n_1 = 878\,829$	$n_2 = 106\,238$
darunter Knaben ...	$m_1 = 454\,597$	$m_2 = 54630$

Das gibt für die Ehelichen bzw. Unehelichen nach Formel (3) die Werte

$$p_1 = 0{,}51728 \qquad p_2 = 0{,}51422.$$

Daraus berechnet sich

$$\mu_1^2 = \frac{p_1 q_1}{n_1+3} = \frac{0{,}51728 \cdot 0{,}48272}{878\,832} = 0{,}00000028417$$

$$\mu_2^2 = \frac{p_2 q_2}{n_2+3} = \frac{0{,}51422 \cdot 0{,}48578}{106\,241} = 0{,}00000235126$$

$$\mu^2 = \mu_1^2 + \mu_2^2 = 0{,}00000263543$$

$$\mu = 0{,}00162,$$

hingegen ist $p_1 - p_2 = 0{,}00306$.

[1]) Statistisches Jahrbuch für das Deutsche Reich 1935, S. 39.

Auf dieser Grundlage allein könnte geurteilt werden, daß ein wesentlicher, d. h. in der Materie liegender Unterschied zwischen der Knabenquote[1]) der ehelich und der der unehelich Geborenen nicht besteht. Jedoch gilt die hier beobachtete Größenbeziehung $p_1 > p_2$ fast durchgängig, so daß es sich um eine weitaus vorwaltende Erscheinung handelt. Der Grund für diese Größenbeziehung liegt in der größeren Fehlgeburtenhäufigkeit beim männlichen Geschlecht und in der größeren Fehlgeburtenhäufigkeit bei unehelichen Geburten.

Zweites Beispiel. Nach derselben Statistik [2]) betrug in dem genannten Jahre die Zahl der

	Lebend-	Tot-
Geborenen	$n_1 = 956\,971$	$n_2 = 28\,096$
darunter Knaben ...	493 473	15 754

Hieraus berechnet sich für die Lebendgeborenen und für die Totgeborenen

$$p_1 = 0{,}51566 \qquad p_2 = 0{,}56072,$$

Daraus berechnet sich die mittlere Abweichung der Differenz $p_2 - p_1$

$$\mu_1^2 = \frac{0{,}51566 \cdot 0{,}48434}{956\,974} = 0{,}0000002610$$

$$\mu_2^2 = \frac{0{,}56072 \cdot 0{,}43928}{28\,099} = 0{,}0000087658$$

$$\mu^2 = \mu_1^2 + \mu_2^2 = 0{,}0000090268$$

$$\mu = 0{,}00300,$$

während $p_2 - p_1 = 0{,}04506$ ist; diese Differenz ist mehr als das 15fache von μ; eine solche Abweichung ist aus zufälligen Störungen nicht zu erwarten, der Unterschied zwischen der Häufigkeit männlicher Geburten bei Lebend- und Totgeborenen erscheint nach diesen Beobachtungen als ein wesentlicher. Zur Gewißheit wird dies, wenn die Beziehung $p_2 > p_1$ zeitlich und örtlich anhält. Nach den vorliegenden Statistiken gilt diese Beziehung durchgängig. Hieraus folgt, daß das männliche Geschlecht vor der Geburt stärker gefährdet ist als das weibliche.

Drittes Beispiel. Aus den Versuchen Darwins über den Einfluß der Abstammung auf das Wachstum der Pflanzen [3]) (Art. 15,3) seien die folgenden Versuchsreihen herausgehoben:

[1]) Die Werte für p_1 bzw. p_2 stimmen in den hier berechneten 5 Dezimalstellen mit den Knabenquoten der ehelich bzw. unehelich Lebendgeborenen infolge des großen n-Wertes überein. Das Entsprechende gilt aus dem gleichen Grunde für die mittlere quadratische Abweichung.

[2]) Statistisches Jahrbuch für das Deutsche Reich 1935, S. 39.

[3]) Vgl. G. U. Yule, Phil. Trans. Roy. Soc., A, vol. 194, 1900, p. 293 u. 295.

	Kreuzung		Selbstbefruchtung	
	Pflanzenhöhe			
	über Mittel	unter Mittel	über Mittel	unter Mittel
Reseda odorata	39	16	25	30
Ipomœa purpurea ...	63	10	18	55

Bei Reseda ist

$$n_1 = n_2 = 55$$

$$p_1 = \frac{40}{57} = 0{,}7018 \qquad p_2 = \frac{26}{57} = 0{,}4561$$

daraus berechnet sich

$$\mu = 0{,}0888,$$

dem steht

$$p_1 - p_2 = 0{,}2457$$

gegenüber.

Bei Ipomœa hat man

$$n_1 = n_2 = 73$$

$$p_1 = \frac{64}{75} = 0{,}8533 \qquad p_2 = \frac{19}{75} = 0{,}2533$$

hiermit wird

$$\mu = 0{,}0643$$

während

$$p_1 - p_2 = 0{,}6000.$$

Ist bei Reseda $p_1 - p_2$ bloß das 2,8fache von μ, kann also auf Grund dieser Versuchsreihe allein eine Begünstigung des Hochwuchses bei Kreuzung gegenüber Selbstbefruchtung in Zweifel gezogen werden, so sprechen bei Ipomœa die Rechnungsergebnisse für eine solche, da hier $p_1 - p_2$ das 9,3fache von μ ausmacht.

§ 2. Die binomiale Verteilung.

111. Die logischen Überlegungen, auf welche sich das bisher über Versuchsreihen Vorgeführte gestützt hat, können weitergeführt und dazu benützt werden, um über die ganze Verteilung solcher Reihen auf apriorischem Wege Aussagen zu gewinnen.

Wir waren davon ausgegangen, daß zwei Ereignisse A, B (Nicht-A) mit den Wahrscheinlichkeiten p, q sich auf N Versuche so verteilen, daß Np Versuche A und Nq Versuche B herbeiführen; die Verteilung wurde als die logisch einzig begründete erkannt.

Diesen Grundgedanken wollen wir jetzt auf den Fall anwenden, daß N Reihen von je n Versuchen über diese Ereignisse ausgeführt werden. Wir numerieren die Versuche jeder Reihe nach ihrer zeitlichen Folge mit 1, 2, ... n, greifen aus jeder Reihe zunächst den ersten Versuch heraus und bezeichnen diese N Versuche als „erste Reihenversuche“. In Weiterführung dieser Überlegung sprechen wir von zweiten, dritten ... $n$ten Reihenversuchen.

Statt immer „logisch begründete“ oder „logisch zu erwartende“ Verteilung zu sagen, soll mit Weglassung dieser Worte kurz von Verteilung gesprochen werden.

Die ersten Reihenversuche geben für A, B die Verteilung Np, Nq.

Bei den zweiten Reihenversuchen ergibt sich bezüglich der Np Fälle für A, B die Verteilung $(Np)p$, $(Np)q$; bezüglich der Nq Fälle die Verteilung $(Nq)p$, $(Nq)q$; im ganzen also ist die Verteilung der Ergebnisse der ersten zwei Reihenversuche durch die Zahlen

$$Np^2,\ 2Npq,\ Nq^2$$

ausgedrückt, entsprechend den Kombinationen, wo zweimal A, je einmal A und B, zweimal B erscheint.

Bei den dritten Reihenversuchen erfolgt die Verteilung von A, B auf die drei eben genannten Kombinationen gemäß den Zahlen

$$(Np^2)p,\ (Np^2)q;\ (2Npq)p,\ (2Npq)q;\ (Nq^2)p,\ (Nq^2)q;$$

im ganzen folgt daraus eine Verteilung der Ergebnisse der ersten drei Reihenversuche nach den Zahlen

$$Np^3,\ 3Np^2q,\ 3Npq^2,\ Nq^3$$

auf die Fälle, wo A dreimal, zweimal, einmal und schließlich gar nicht vorkommt.

Die so gefundenen Verteilungszahlen, jedesmal zu einer Summe vereinigt, ergeben also der Reihe nach

bei den ersten Reihenversuchen... $Np + Nq = N(p+q)$

bei den zweiten Reihenversuchen.. $Np^2 + 2Npq + Nq^2 = N(p+q)^2$

bei den dritten Reihenversuchen... $Np^3 + 3Np^2q + 3Npq^2 + Nq^3 = N(p+q)^3$;

bei den $n$ten Reihenversuchen.... $Np^n + N\frac{n}{1}p^{n-1}q +$

$$+ N\frac{n(n-1)}{1\cdot 2}p^{n-2}q^2 + \cdots + N\frac{n(n-1)\ldots(n-m+1)}{1\cdot 2\ldots m}p^{n-m}q^m + \cdots + Nq^n = N(p+q)^n$$

Führt man Nmal je n Versuche über zwei entgegengesetzte Ereignisse A, B mit den festbleibenden Wahrscheinlichkeiten p, q aus, so ist in den Reihenergebnissen aus logischen Gründen eine solche Verteilung zu erwarten, wie die durch die Glieder der Entwicklung $N(p+q)^n$, also durch

$$\left.\begin{aligned} Np^n + N\frac{n}{1}p^{n-1}q + N\frac{n(n-1)}{1\cdot 2}p^{n-2}q^2 + \ldots \\ + N\frac{n(n-1)\ldots(n-m+1)}{1\cdot 2\ldots m}p^{n-m}q^m + \cdots + Nq^n \end{aligned}\right\} \qquad (1a)$$

beschrieben ist in dem Sinne, daß in $N\frac{n(n-1)\ldots(n-m+1)}{1\cdot 2\ldots m}p^{n-m}q^m$ Reihen von den N Reihen je $(n-m)$mal A und $m$mal B in was immer für einer Anordnung erscheint.

Unterdrückt man den Faktor N, so bleiben die Häufigkeiten der verschiedenen Ergebnisse oder Erfolge, gerechnet auf eine Reihe, also die Wahrscheinlichkeiten dieser Erfolge zurück.

Nach (1a) stellt sich die Wahrscheinlichkeit P_m dafür, daß bei n Versuchen $(n-m)$ mal das Ereignis A und m mal das Ereignis B eintritt, auf

$$P_m = \frac{n!}{(n-m)!\,m!}\,p^{n-m}\,q^m = \binom{n}{m}\,p^{n-m}\,q^m. \tag{1b}$$

Die durch die Formel (1b), die man vielfach die Newtonsche Formel nennt, bestimmte Verteilung bezeichnet man als Bernoullische[1]) Verteilung.

Die Koeffizienten

$$1,\ \frac{n}{1},\ \frac{n(n-1)}{1\cdot 2},\ \ldots 1 \tag{1c}$$

bilden eine symmetrische, von den Enden gegen die Mitte ansteigende Zahlenfolge mit einem oder zwei größten Gliedern, je nachdem n gerade oder ungerade ist[2]).

Die Potenzprodukte

$$p^n,\ p^{n-1}q,\ p^{n-2}q^2,\ .\ .\ q^n \tag{1d}$$

bilden eine monotone, u. zw. eine fallende geometrische Reihe, wenn $p > q$, eine steigende, wenn $p < q$, eine weder steigende noch fallende, wenn $p = q = \frac{1}{2}$ ist.

Die Produkte aus homologen Gliedern der beiden Reihen (1a) und (1b), also eben die relativen Häufigkeiten der Reihenergebnisse, werden hiernach in den Fällen $p \mp q$ eine asymmetrische Zahlenfolge bilden, im allgemeinen mit einem größten Gliede. Um die Lage dieses Gliedes zu bestimmen, bilden wir den Quotienten aus dem $(m+1)$ten Gliede durch das $m$te, dieser ist

$$\frac{n-m+1}{m}\,\frac{q}{p};$$

das Ansteigen dauert so lange als

$$\frac{n-m+1}{m}\,\frac{q}{p} > 1,$$

also als

$$m < nq + q;$$

das Abfallen so lange als

$$\frac{n-m+1}{m}\,\frac{q}{p} < 1,$$

also als

$$m > nq + q;$$

[1]) Jacob Bernoulli, Wahrscheinlichkeitsrechnung (Ars conjectandi, Basel 1713). Ostwalds Klassiker der exakten Wissenschaften Nr. 107 und 108.

[2]) Die Bionomialkoeffizienten spielen in der Theorie der statistischen Erhebung eine Rolle. Vgl. hierzu F. Rusam, Systematik der statistischen Erhebung. Neumanns Zeitschrift für Versicherungswesen. Berlin 1937, Nr. 3, S. 53 u. f.

somit gehört die größte unter $nq+q$ liegende ganze Zahl dem größten Gliede als Exponent von q an. Anderseits dauert das Ansteigen so lange als

$$n-m > np-q,$$

das Abfallen so lange als

$$n-m < np-q;$$

sonach ist der Exponent von p im größten Gliede die kleinste über $np-q$ liegende ganze Zahl.

In dem Falle, wo $nq+q$ selbst eine ganze Zahl ist, ist es auch $np-q$; dann geht das Ansteigen bis $m=nq-p$, bei $m+1=nq+q$ besteht Gleichheit mit dem vorangehenden Gliede, und bei $m+2=nq+2-p$ beginnt das Abfallen.

Beispielsweise ist bei $p=\frac{2}{5}$, $q=\frac{3}{5}$ das Exponentenpaar $n-m$, m des maximalen Gliedes für die folgenden n Werte:

n	$n-m$	m
100	40	60
101	40	61
102	41	61
103	41	62
104	{ 42	62
	{ 41	63

Für $n=104$ herrscht bis zum Glied mit dem ersten Exponentenpaar 42, 62 Ansteigen, das nächste Glied mit dem zweiten Paar ist dem Glied mit dem ersten Paar gleich und bei dem Glied mit dem Exponentenpaar 40, 64 beginnt das Abfallen.

Das Verhältnis der Exponenten $n-m$, m von p und q im maximalen Gliede kommt hiernach dem Verhältnis der Wahrscheinlichkeiten p, q entweder gleich oder näher als in jedem andern Gliede der Entwicklung. Daraus folgt, daß die Längen der beiden Reihenabschnitte, die durch das maximale Glied oder durch die Zäsur zwischen den beiden größten Gliedern (in dem Ausnahmefalle) gebildet werden, sich verhalten wie q zu p; folglich besteht bei $p>q$ linksseitige, bei $p<q$ rechtsseitige Asymmetrie. Ihr Ausmaß hängt von dem Verhältnis $p:q$ und von n ab.

112. Über die Ausrechnung einer Verteilung für besondere Werte von p, q, n ist, so lange n einen mäßigen Wert hat, nichts zu bemerken. Bei größeren Werten aber empfiehlt es sich, mit der Ausrechnung des maximalen Gliedes zu beginnen und von diesem aus mit Hilfe der Quotienten benachbarter Glieder die übrigen Glieder zu beiden Seiten des maximalen abzuleiten. Einige Beispiele werden dies am besten klar machen.

1) Die unmittelbare Ausrechnung von 10000 $(0{,}1+0{,}9)^2$ gibt die Verteilung: 100, 1800, 8100; wenn also über zwei Ereignisse A, B mit den Wahrscheinlichkeiten 0,1, 0,9 10000 Versuchspaare angestellt werden, so sind

Entwicklung von $10000\,(0{,}4+0{,}6)^{20}$.

m	z
0	—
1	—
2	—
3	—
4	3
5	13
6	49
7	146
8	355
9	710
10	1171
11	1597
12	1797
13	1659
14	1244
15	746
16	350
17	124
18	31
19	5
20	—
	10000

nur 100 Paare zu erwarten, die aus A allein, 1800 Paare, die aus A und B, und 8100 Paare, die aus B allein bestehen. Die Verteilung ist eine einseitige, weil das maximale Glied mit einem Endglied zusammenfällt.

2) Die unmittelbare Ausführung von $10000\,(0{,}4+0{,}6)^2$ ergibt die Verteilung 1600, 4800, 3600, die bereits zweiseitig und dabei asymmetrisch ist.

3) Wenn es sich um die Entwicklung von $10000\,(0{,}4+0{,}6)^{20}$ handelt, wäre das direkte Verfahren schon mühsam. Man rechnet zuerst das maximale Glied

$$10000\cdot\frac{20!}{8!\,12!}\left(\frac{4}{10}\right)^8\left(\frac{6}{10}\right)^{12}=\frac{13\cdot 14\ldots 20}{1\cdot 2\ldots 8}\,\frac{4^8\;6^{12}}{10^{16}}=1797$$

aus und leitet daraus mit Hilfe der Quotienten

$$\frac{12}{9}\cdot\frac{4}{6},\;\frac{11}{10}\cdot\frac{4}{6},\;\frac{10}{11}\cdot\frac{4}{6},\;\ldots\;\frac{1}{20}\cdot\frac{4}{6}$$

und

$$\frac{8}{13}\cdot\frac{6}{4},\;\frac{7}{14}\cdot\frac{6}{4},\;\frac{6}{15}\cdot\frac{6}{4},\;\ldots\;\frac{1}{20}\cdot\frac{6}{4}$$

die Glieder links und rechts von ihm ab; man erhält so die nebenstehende, vertikal mit dem linken Ende oben angeschriebene Tabelle.

Die Verteilung zeigt schon einen beträchtlichen Grad von Symmetrie. Ergebnisreihen, in welchen A gar nicht und solche, in welchen B weniger als viermal vorkommt, sind bei 10000 Versuchsserien nicht zu erwarten.

Mit wachsendem n nimmt unter sonst gleichbleibenden Umständen der besetzte Teil der Verteilungstafel an Ausdehnung zu, während gleichzeitig eine Verflachung, d. h. eine Abnahme der Häufigkeitszahlen, im Kernstück eintritt. Das zeigt die weiter mitgeteilte Ausrechnung von $10000\left(\frac{2}{3}+\frac{1}{3}\right)^{42}$. Zur besseren Übersicht ist die Tabelle so angeordnet, daß das Maximalglied hervortritt und die zu ihm symmetrisch liegenden Glieder nebeneinander stehen; der hohe Grad der Symmetrie ist augenfällig sowohl in der beiderseitigen Ausdehnung als auch in der Größe der Zahlen. Reihen, in welchen das Ereignis B weniger als 3mal und mehr als 26mal auftritt, kommen in 10000 Serien nicht vor. Die Summe der Häufigkeitszahlen beträgt 10003 statt 10000 eine Folge der Aufrundung auf ganze Zahlen. (S. Tab. auf S. 271.)

113. Wenn der Umfang n der Reihen sehr groß ist, wird selbst die genaue Auswertung des Maximalgliedes eine beschwerliche, kaum zu bewältigende Arbeit. Man benützt dann eine Näherungsformel, die sich ergibt, wenn man in dem Aus-

druck des Maximalgliedes für die darin vorkommenden Faktoriellen Näherungen nach der Stirlingschen Formel

$$x! = x^{x+\frac{1}{2}} e^{-x} \sqrt{2\pi} \qquad (2)$$

einsetzt, von der auch dann Gebrauch gemacht werden kann, wenn x nicht ganzzahlig ist, was sonst in dem linksstehenden Symbol vorausgesetzt wird. Werden die Exponenten von p, q in dem maximalen Gliede durch die ihnen jedenfalls sehr nahe liegenden Zahlen np, nq ersetzt, so schreibt sich dieses Glied

$$\frac{n!}{(np)!\,(nq)!}\, p^{np} q^{nq}. \qquad (3)$$

Setzt man darin für $n!$, $(np)!$, $(nq)!$ die Ausdrücke nach der Stirlingschen Formel ein, so ergibt sich nach gehöriger Kürzung für das Maximalglied die Näherungsformel

$$\frac{1}{\sqrt{2\pi npq}} \qquad (4)$$

Von dieser Formel soll Gebrauch gemacht werden, um an einem ausgerechneten Beispiel zu zeigen, daß selbst bei beträchtlichem Größenunterschied zwischen p und q wenigstens in dem Kernstück der Verteilung ein hoher Grad von Symmetrie besteht, wenn nur n genügend groß ist.

Entwicklung von $10000\left(\frac{2}{3}+\frac{1}{3}\right)^{42}$

m	z	z	m
.	.	—	42
.	.	—	41
.	.	—	40
.	.	—	39
.	.	—	38
.	.	—	37
.	.	—	36
.	.	—	35
.	.	—	34
.	.	—	33
.	.	—	32
.	.	—	31
.	.	—	30
.	.	—	29
0	—	—	28
1	—	—	27
2	—	1	26
3	1	3	25
4	3	8	24
5	11	21	23
6	33	49	22
7	85	103	21
8	185	197	20
9	350	343	19
10	578	543	18
11	840	782	17
12	1085	1022	16
13	1252	1211	15
14	1297	.	.

4) Es soll das Mittelstück der Verteilung ausgerechnet werden, die zusammenfassend durch $100000\,(0{,}1 + 0{,}9)^{1000}$ dargestellt ist.

Das Maximalglied hat nach der Formel (4) den Wert

$$\frac{100000}{\sqrt{180\pi}} = 4205{,}2\,;$$

von diesem geht die Berechnung der anderen Glieder nach der linken Seite mittels der Quotienten

$$\frac{900}{101}\cdot\frac{1}{9},\quad \frac{899}{102}\cdot\frac{1}{9},\quad \frac{898}{103}\cdot\frac{1}{9},\quad \ldots\quad \frac{1}{1000}\cdot\frac{1}{9},$$

nach der rechten Seite mittels der Quotienten

$$\frac{100}{901}\cdot\frac{9}{1},\quad \frac{99}{902}\cdot\frac{9}{1},\quad \frac{98}{903}\cdot\frac{9}{1},\quad \ldots\quad \frac{1}{1000}\cdot\frac{9}{1}.$$

Kernstück der Entwicklung von 100000 $(0{,}1 + 0{,}9)^{1000}$.

m	z	z	m
883	849	841	917
884	1011	1020	916
885	1193	1221	915
886	1394	1444	914
887	1613	1685	913
888	1847	1942	912
889	2094	2211	911
890	2350	2487	910
891	2611	2763	909
892	2871	3033	908
893	3125	3290	907
894	3366	3527	906
895	3588	3737	905
896	3784	3914	904
897	3949	4053	903
898	4077	4149	902
899	4163	4200	901
900	4205		

Die nebenstehende Tabelle gibt bloß die 35 Glieder, deren mittelstes das Maximalglied ist.

Die Anordnung ist so getroffen wie bei der vorangehenden Tabelle, um die Vergleichung gleichgestellter Glieder links und rechts vom maximalen zu erleichtern. Hervorzuheben ist der rasche Abfall und der Umstand, daß die Summe dieser 35 angeschriebenen Glieder 93607 beträgt, so daß auf die 966 übrigen Glieder nur die Summe 6393 entfällt.

An die Newtonsche Formel (1b in Art. 111) sei unter Hinweis auf Riebesell[1]) noch folgende Bemerkung angeschlossen. Wir denken uns aus einer Urne, die N Kugeln, und zwar Np weiße und $N(1-p) = Nq$ schwarze Kugeln enthält, n Ziehungen vorgenommen. Bei jeder Ziehung nimmt man eine Kugel heraus und legt nach der Ziehung $(1+\varepsilon)$ Kugeln derselben Farbe wieder hinein. Wir setzen $\frac{\varepsilon}{N} = \delta$. Bei der Newtonschen Formel ist $\varepsilon = 0$. Für $\varepsilon > 0$ gelangt man zu der folgenden Formel

$$P_m = \frac{\quad}{m!\,(n-m)!}$$

$$\cdot \frac{p\,(p+\delta)\,(p+2\,\delta)\ldots[p+(n-m-1)\,\delta]\cdot q\,(q+\delta)\ldots[q+(m-1)\,\delta]}{1\,(1+\delta)\,(1+2\,\delta)\ldots[1+(n-1)\,\delta]}$$

Aus dieser Gleichung erhält man, wenn $n \to \infty$

$$\frac{1}{P_m}\,\frac{d\,P_m}{dm} = \frac{(1-2\,\delta)\,m-(q-\delta)}{m\,\delta\,(1-m)}$$

Diese Gleichung, die die „Wahrscheinlichkeitsansteckung“ kennzeichnet, wird bei statistischen Untersuchungen über „verkettete“ Vorgänge angewandt, z. B. in der Todesursachenstatistik bei der schärferen Erfassung ansteckender Krankheiten. Die Wahrscheinlichkeitsansteckung ist auch für erblichkeitsstatistische Forschungen von Bedeutung.[2])

[1]) P. Riebesell, Einführung in die Sachversicherungs-Mathematik. Veröffentlichungen des Deutschen Vereins für Versicherungs-Wissenschaft, Heft 56, Berlin 1936, S. 27 u. f.

[2]) Vgl. F. Eggenberger und G. Polya, Über die Statistik verketteter Vorgänge. Zeitschrift für angewandte Mathematik und Mechanik 1923, S. 279 u. f.

114. Man kann die binomiale Verteilung ebenso wie jede andere durch ein Häufigkeitspolygon darstellen. Es gibt ein einfaches Konstruktionsverfahren, um die Häufigkeitspolygone zu $N(p+q)^n$ für die aufeinanderfolgenden Werte von n bei gleichbleibendem p, q, N herzustellen; dasselbe gründet sich auf folgende Bemerkungen. Aus den beiden Nachbargliedern

$$u_{r-1}^{(n)} = \binom{n}{r-1} p^{n-r+1} q^{r-1}$$

$$u_r^{(n)} = \binom{n}{r} p^{n-r} q^r$$

Fig. 33. Zur Konstruktion des Häufigkeitspolygons zu $N(p+q)^n$.

ergibt sich nach dem Schema

$$q\,u_{r-1}^{(n)} + p\,u_r^{(n)} = \left[\binom{n}{r-1} + \binom{n}{r}\right] p^{n+1-r} q^r = \binom{n+1}{r} p^{n+1-r} q^r = u_r^{(n+1)}$$

ein Glied der Entwicklung von $N(p+q)^{n+1}$. Konstruktiv kann dies so durchgeführt werden, daß man auf zwei Parallelen, deren Abstand als Einheit in der horizontalen Achse gilt, $u_{r-1}^{(n)}$ und $u_r^{(n)}$ nach irgend einem Maßstabe abträgt, die Endpunkte B, D, Fig. 33, verbindet, dann eine dritte Parallele zieht, welche die Basis AC in die Teile p, q zerlegt; dann ist

$$EF = FG + GE = q \cdot AB + p \cdot CD = q u_{r-1}^{(n)} + p\, u_r^{(n)} = u_r^{(n+1)}$$

Wie hiervon Gebrauch gemacht werden kann, soll an der Verteilung 1000 $\left(\frac{1}{4} + \frac{3}{4}\right)^n$ gezeigt werden. Auf der horizontalen Achse der Fig. 34 ist die Einheit wiederholt aufgetragen und in den Punkten 0, 1, 2, ... sind Vertikallinien als die Träger der Eckpunkte der Polygone gezogen; die punktierten Vertikalen teilen die zugehörigen Einheitsstrecken im Verhältnis von 1 : 3. Das Polygon zu 1000 $\left(\frac{1}{4} + \frac{3}{4}\right)^1$ hat außer dem Anfangspunkt 0 und dem Endpunkt 3 zwei Ecken in den Höhen 250, 750, die auf dem links errichteten Häufigkeitsmaßstab abzulesen sind.

Die Punkte, in welchen die Seiten dieses ersten Polygons die punktierten Vertikalen treffen, projiziere man horizontal auf die nachfolgenden Hauptvertikalen und erhält so die Ecken des Polygons zu $1000\left(\frac{1}{4} + \frac{3}{4}\right)^2$, zu welchen 0 und 4 als Endpunkte hinzukommen.

Durch die Punkte, in welchen die Seiten dieses zweiten Polygons die punktierten Vertikalen schneiden, führe man wieder Horizontale bis an die nachfolgenden Hauptvertikalen und erhält so die Eckpunkte des Polygons zu 1000 $\left(\frac{1}{4} + \frac{3}{4}\right)^3$, zu welchem 0, 5 als Endpunkte gehören.

Das Verfahren ist fortgesetzt bis 1000 $\left(\frac{1}{4} + \frac{3}{4}\right)^5$; die Polygone sind durch die Bezifferungen (1), (1); (2), (2); ... (5), (5) genügend kenntlich gemacht. In jedem Polygon ist die Summe der Eckenhöhen = 1000, die Flächen der Polygone sind von gleicher Größe, nämlich 1000 Flächeneinheiten. Man beobachte die fort-

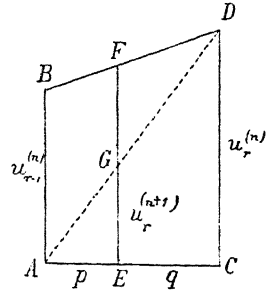
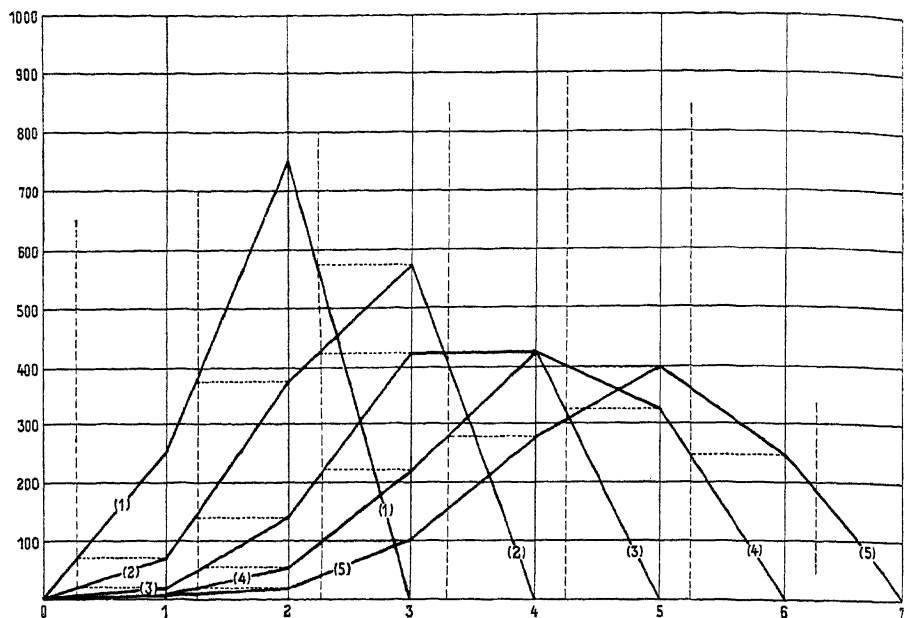



Fig. 34. Häufigkeitspolygone zu den binomialen Verteilungen $1000\left(\frac{1}{4}+\frac{3}{4}\right)^n$.

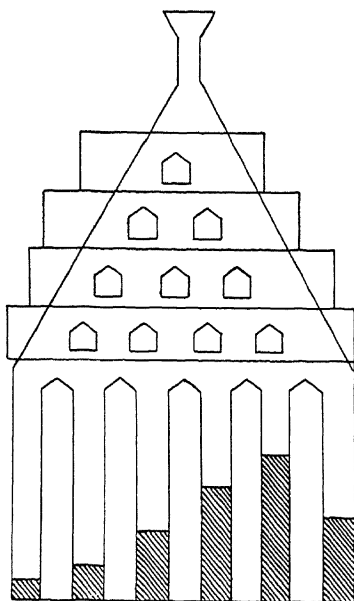



Fig. 35. Pearsons Binomialapparat.

schreitende Streckung und Verflachung der Polygone; auch die zunehmende Symmetrie macht sich in der Figur bemerkbar. Weitgehende Fortsetzung der Konstruktion scheitert an der Häufung der Linien und Fehler und an der wachsenden horizontalen Ausdehnung der Zeichnung.

115. Galton hat einen Apparat konstruiert und Pearson[1]) ihn verbessert, durch welchen der experimentelle Nachweis der Entstehung einer binomialen Verteilung gegeben werden soll. In der verbesserten Form besteht der Apparat, Fig. 35, in einem flachen Kasten, dessen Vorderwand aus Glas ist; auf der hölzernen Rückwand sind auf Schlitten gleiche und in gleichen Abständen von-

[1]) K. Pearson, Contributions to the Mathematical Theory of Evolution. Phil. Trans. Roy. Soc., A, Vol. 186, 1895, p. 345 und Plate 7, Fig. 2.

einander angebrachte Keile verschiebbar angeordnet; die untersten dieser Keile reichen bis zum Boden des Kastens und teilen seinen untern Teil in eine Anzahl gleicher Fächer. Oben befindet sich ein Trichter zum Einfüllen eines körnigen Stoffes, z. B. kleiner Schrotkörner, irgendwelcher Samenkörner o. dgl. Beim Auftreffen der Körner auf einen solchen Keil erfolgt eine Teilung ihres Stromes nach links und rechts, und es hängt von der Einstellung der Keilreihen gegeneinander ab, wieviel in den linken und wieviel in den rechten Zwischenraum der darunter befindlichen Keilpaare abfließt; diese Teilung wiederholt sich immer wieder, und zwar immer in demselben Verhältnis, so oft der Strom einen neuen Keil trifft, sofern nur die Einstellung durchwegs das gleiche Verhältnis einhält. Der Enderfolg ist der, daß sich die unteren Fächer in verschiedenem Maße mit Körnern füllen, und zwar geschieht dies offenkundig in der dem Einstellungsverhältnis entsprechenden binomialen Verteilung. Das Beschicken des Apparates, den man gelegentlich auch Binomialmaschine genannt hat, geschieht bei geneigter Rückwand.

116. Die Frage nach dem Mittelwerte der Wiederholungszahl eines der beiden Ereignisse, z. B. des Ereignisses A von der Wahrscheinlichkeit p, die durch apriorische Erwägung erschlossen, nicht durch Beobachtung bestimmt wird, erledigt sich dadurch, daß man die Summe der Produkte der möglichen Wiederholungszahlen mit den zugehörigen Wahrscheinlichkeiten bildet, also der Produkte folgender Wertepaare:

Wahrscheinlichkeit	Wiederholungszahl
q^n	0
$n q^{n-1} p$	1
$\frac{n(n-1)}{1 \cdot 2} q^{n-2} p^2$	2
p^n	n;

die Summe aus diesen Produkten ist das gesuchte arithmetische Mittel, also

$$\left.\begin{aligned} M &= np\left[q^{n-1} + (n-1)q^{n-2}p + \frac{(n-1)(n-2)}{1 \cdot 2} q^{n-3} p^2 + \cdots + p^{n-1}\right] \\ &= np\,(q+p)^{n-1} = np. \end{aligned}\right\} \quad (5)$$

In derselben Weise erhält man die mittlere Wiederholungszahl des Ereignisses B von der Wahrscheinlichkeit q gleich nq.

Führt man also N Versuchsreihen von je n Gliedern aus und zählt in jeder das Auftreten von A, nennt allgemein seine Wiederholungszahlen $\nu_1, \nu_2, \ldots \nu_N$ in der ersten, zweiten, ... N-ten Versuchsreihe, so kommt

$$\frac{\nu_1 + \nu_2 + \cdots + \nu_N}{N}$$

im allgemeinen um so näher dem a priori bestimmten Werte np, je größer die Reihenzahl N ist.

Als Streuungsmaß der Reihe der Einzelwerte $\nu_1, \nu_2, \ldots \nu_N$ benützen wir wie üblich die mittlere Abweichung. Um sie zu erhalten, stützen wir uns auf das in Art. 59 erörterte Verfahren und multiplizieren die Wahrscheinlichkeiten mit den Quadraten der zugehörigen Wiederholungszahlen, also in diesem Falle die eben verwendeten Produkte nochmals mit den rechts angeschriebenen Zahlen; die Summe der neuen Produkte, vermindert um das Quadrat von np, gibt das Quadrat der mittleren Abweichung; also ist

$$\mu^2 = np\left[q^{n-1} + 2(n-1)\,q^{n-2}p + 3\frac{(n-1)(n-2)}{1\cdot 2}q^{n-3}p^2 + \ldots + \; p^{n-1}\right] - n^2p^2.$$

In der Klammer stehen die Glieder der Entwicklung von $(q+p)^{n-1}$, multipliziert mit den Zahlen $1, 2, 3, \ldots n$; oben ist aber gefunden worden, daß die Glieder der Entwicklung von $(q+p)^n$, multpliziert mit den Zahlen $0, 1, 2, \ldots n$, in der Summe np geben; infolgedessen ist die Klammersumme $(n-1)p+1$, da die Summe der einfachen Glieder von $(q+p)^{n-1}$ gleich 1 ist, demnach ist schließlich

$$\mu^2 = np\,[(n-1)p+1] - n^2p^2 = np - np^2 = npq. \qquad (6)$$

Wir sind also zu demselben Resultat gekommen, das sich in Art. 106 durch eine andere Schlußweise ergeben hat. Wir sind gegenüber den dortigen Ausführungen insofern hinausgekommen, als wir jetzt die a priori zu erwartende Verteilung der N Beobachtungsreihen nach den Wiederholungszahlen der sie zusammensetzenden Ereignisse kennen und sie mit der wirklich eingetretenen Verteilung vergleichen können. Wenn bei den Versuchen alle die Voraussetzungen erfüllt sind, von welchen dort die Rede war, so sind die Abweichungen der a priori berechneten Verteilung von der beobachteten bloße Wirkung des Zufalls. Ob dem so ist, kann außer aus der Verteilung selbst auch aus der mittleren Abweichung beurteilt werden, die sich aus ihr ergibt, indem man sie mit der a priori nach der Formel $\mu = \sqrt{npq}$ berechneten vergleicht; doch kann das Urteil nie ein endgültiges sein.

Als Beispiel einer solchen Vergleichung benützen wir die beiden Würfelversuchsreihen (Art. 105, 1). Für die erste besteht die theoretische Verteilung in der Entwicklung von $4096\left(\frac{1}{2}+\frac{1}{2}\right)^{12}$, d. i. in den Gliedern von $(1+1)^{12}$; für die andere in den Gliedern der Entwicklung von $6500\left(\frac{1}{2}+\frac{1}{2}\right)^{12}$.

Nebenstehend sind diese theoretischen Verteilungen mit den beobachteten zusammengestellt und an letztere die Nebenrechnungen zur Bestimmung von M und μ angeschlossen.

Die theoretischen Resultate sind für beide Fälle die gleichen, nämlich:

$$M = 12\cdot\frac{1}{2} = 6$$

$$\mu = \sqrt{12\cdot\frac{1}{2}\cdot\frac{1}{2}} = \sqrt{3} = 1{,}732.$$

Die Übereinstimmung ist befriedigend, in der umfangreicheren Reihe besser als in der minder umfangreichen. Es sei nochmals betont, daß die Berechnungen dieses Artikels unter der Voraussetzung stehen, daß die Wahrscheinlichkeit p apriorisch erschlossen, nicht empirisch bestimmt ist.

Erfolge	Häufigkeit berechnet	Häufigkeit beobachtet		
0	1	.	.	.
1	12	7	.	.
2	66	60	4089	16960
3	220	198	4029	32602
4	495	430	3831	12931
5	792	731	3401	9100
6	924	948	2670	5699
7	792	847	1722	3029
8	495	536	875	1307
9	220	257	339	432
10	66	71	82	93
11	12	11	11	11
12	1	.	.	.
	4096	4096		

$$M = 1 + \frac{4089 + 16960}{4096} = 6{,}139$$

$$\mu^2 = \frac{4089 + 3 \cdot 16960 + 2 \cdot 32602}{4096} - 5{,}139^2 = 2{,}9298$$

$$\mu = 1{,}712.$$

Erfolge	Häufigkeit berechnet	Häufigkeit beobachtet		
0	2	1	.	.
1	19	14	6499	33257
2	105	103	6485	78190
3	349	302	6382	26772
4	786	711	6080	20390
5	1257	1231	5369	14310
6	1466	1411	4138	8941
7	1257	1351	2727	4803
8	786	844	1376	2076
9	349	391	532	700
10	105	117	141	168
11	19	21	24	27
12	2	3	3	3
	6502	6500		

$$M = \frac{6499 + 33257}{6500} = 6{,}116$$

$$\mu^2 = \frac{6499 + 3 \cdot 33257 + 2 \cdot 78190}{6500} - 6{,}116^2 = 3{,}0022$$

$$\mu = 1{,}733.$$

Bei den vorstehenden Betrachtungen ist die Reihenfolge, in der die beiden Ereignisse auftreten, nicht in Betracht gezogen worden. Dies hat seinen Grund darin, daß es für die Beurteilung der relativen Häufigkeit des Auftretens eines Ereignisses gleichgültig ist, in welcher Reihenfolge die beiden Ereignisse A und B aufeinander folgen, d, h. ob in einzelnen Teilen der Versuchsreihe immer auf das Ereignis A auch sofort wieder das gleiche Ereignis A folgt und in anderen Teilen der Versuchsreihe dasselbe vom zweiten Ereignis B gilt, oder ob auf das Ereignis A immer das Ereignis B und umgekehrt auf das Ereignis B das Ereignis A folgt. Tritt in der Versuchsreihe das Ereignis A l-mal ununterbrochen hintereinander auf, so bezeichnet man nach L. v. Bortkiewicz[1]) diese Folge von l Ereignissen A als Iteration von der Länge l.

Das Problem der Iterationen hat eine mathematische und eine statistische Seite. Die mathematische Seite besteht darin zu berechnen, wie oft in einer bestimmten Versuchsreihe eine Iteration von der Länge l zu erwarten ist, und auf statistischem Wege kann die Häufigkeit des Auftretens einer Iteration von der Länge l empirisch bestimmt werden. Es gilt nun, die wahrscheinlichkeitstheoretisch berechneten Häufigkeiten mit den empirisch bestimmten Häufigkeiten zu vergleichen. Bei den Geburteneintragungen liegt die Übereinstimmung im wesentlichen vor.[2])

§ 3. Die normale Häufigkeitskurve.

117. Anknüpfend an die letzten Erörterungen muß gesagt werden, daß die Anwendbarkeit der binomialen Verteilung nicht weitreichend ist. Selbst wenn man davon absieht, daß sie naturgemäß nur auf Materien paßt, die sich mit Ziehungen, Würfelversuchen u. ä. vergleichen lassen, bleibt doch die Umständlichkeit der Rechnungen bestehen.

Daher wird es von Vorteil sein, die binomiale Verteilung durch eine stetige, geometrisch gesprochen, ihr Häufigkeitspolygon durch eine Kurve zu ersetzen. Der Weg, der sich dazu darbietet, ist die Aufsuchung der Grenze, welcher sich die binomiale Verteilung nähert, wenn die Zahl n unbegrenzt wächst.

Zum Ausgangspunkt werde das Maximalglied genommen, das in der Form

$$z_0 = \frac{n!}{(np)!\,(nq)!}\,p^{np}q^{nq} \tag{1}$$

geschrieben werden kann (Art. 113, (3)); das eine der Glieder, die um x Plätze von ihm entfernt sind, hat den Ausdruck

$$\frac{n!}{(np+x)!\,(nq-x)!}\,p^{np+x}q^{nq-x}. \tag{2}$$

[1]) L. v. Bortkiewicz, Die Iterationen. Ein Beitrag zur Wahrscheinlichkeitstheorie. Berlin 1917, S. 21 u. f. Vgl. auch F. Böhm, Grundfragen der angewandten Wahrscheinlichkeitsrechnung und der theoretischen Statistik, insbesondere das Problem der reinen Gruppen. Archiv für mathematische Wirtschafts- und Sozialforschung. 1936, Heft 1, S. 17 u. f. und Heft 2, S. 69 u. f.

[2]) Vgl. F. Böhm, Grundfragen der angewandten Wahrscheinlichkeitsrechnung und der theoretischen Statistik, insbesondere das Problem der reinen Gruppen. Archiv für mathematische Wirtschafts- und Sozialforschung. 1936, Heft 1, S. 23 u. f.

darin sind, wenn x jede positive wie negative ganze Zahl bedeuten kann, alle Glieder der Entwicklung inbegriffen. Die Wahrscheinlichkeit p wird hierbei wiederum zunächst als logisch erschlossen, nicht empirisch bestimmt, vorausgesetzt.

Die Bezeichnung durch den Buchstaben z soll der Vorstellung Rechnung tragen, daß es sich um die Aufsuchung einer stetigen Funktion oder um die Ordinate einer Kurve handelt; dazu ist aber erforderlich, von der bisher festgehaltenen Vorstellung eines ganzzahligen x abzugehen und sich darunter eine stetige reelle Variable zu denken.

Die Natur der Verteilung hängt nicht von den absoluten Werten der z, sondern bloß von deren Verhältnissen ab; wir führen daher die weiteren Rechnungen an dem Quotienten

$$\frac{z}{z_0} = \frac{(np)!\,(nq)!}{(np+x)!\,(nq-x)!}\left(\frac{p}{q}\right)^x$$

aus und beginnen damit, daß wir die Fakultäten, von welchen wir voraussetzen, daß sie insgesamt zu großen Zahlen gehören, nach der Stirlingschen Formel (Art. 113, (2)) ausdrücken; das gibt nach entsprechender Kürzung und einigen Umformungen in erster Näherung

$$\frac{z}{z_0} = \frac{1}{\left(1+\frac{x}{np}\right)^{np+x+\frac{1}{2}}\left(1-\frac{x}{nq}\right)^{nq-x+\frac{1}{2}}}$$

Der natürliche Logarithmus des Nenners beträgt

$$\left(np+x+\frac{1}{2}\right) ln\left(1+\frac{x}{np}\right)+\left(nq-x+\frac{1}{2}\right) ln\left(1-\frac{x}{nq}\right).$$

Wenn x so eingeschränkt wird, daß $\frac{x}{np}$, $\frac{x}{nq}$ echte Brüche bleiben, so läßt sich $ln\left(1+\frac{x}{np}\right)$ bzw. $ln\left(1-\frac{x}{nq}\right)$ nach der allgemeinen Formel

$$ln\,(1+y) = y-\frac{y^2}{2}+\frac{y^3}{3}\mp\cdots,$$

wobei $|y|<1$ ist, in eine Reihe entwickeln. Man erhält auf diese Weise für den Logarithmus des Nenners folgenden Ausdruck:

$$\left(np+x+\frac{1}{2}\right)\left(\frac{x}{np}-\frac{x^2}{2n^2p^2}\pm\cdots\right)-\left(nq-x+\frac{1}{2}\right)\left(\frac{x}{nq}+\frac{x^2}{2n^2q^2}+\cdots\right);$$

bis auf Glieder zweiten Grades in x entwickelt, gibt dies:

$$\frac{(q-p)x}{2npq}+\frac{x^2}{2npq}-\frac{x^2}{4n^2p^2}-\frac{x^2}{4n^2q^2}.$$

Nun genügt es, mit x nicht über ein mäßiges Vielfaches der mittleren Abweichung $\sqrt{npq}$ zu gehen, wie noch später erklärt werden soll, nicht über das Drei- bis Vierfache; mit andern Worten, x braucht nur bis zu Werten von der

Größenordnung $\sqrt{n}$ geführt zu werden; unter solcher Annahme ist das zweite Glied das vorwaltende, die beiden letzten Glieder sollen, da sie gegenüber dem ersten und zweiten Gliede sehr klein sind, unterdrückt werden.

Mit diesem Grade der Näherung kann der Nenner von $\frac{z}{z_0}$ durch

$$e^{\frac{x^2}{2npq}+\frac{(q-p)x}{2npq}}$$

ersetzt werden, so daß nunmehr

$$\frac{z}{z_0}=e^{-\frac{x^2}{2npq}+\frac{(p-q)x}{2npq}}$$

Entwickelt man $e^{\frac{(p-q)x}{2npq}}$ nach der allgemeinen Formel

$$e^y=1+\frac{y}{1!}+\frac{y^2}{2!}+\cdots$$

wobei $-\infty<y<+\infty$ ist, in eine Reihe, so erhält man mit demselben Grade der Approximation

$$\frac{z}{z_0}=e^{-\frac{x^2}{2npq}}\left(1+\frac{(p-q)x}{2npq}+\cdots\right). \tag{3}$$

Die Exponentialfunktion, die in diesem Ausdruck auftritt, ist eine symmetrische (weil gerade) Funktion; durch den Klammerausdruck wird Asymmetrie herbeigeführt, und zwar wird bei $p>q$ durch positive x eine Vergrößerung, durch negative x eine Verminderung bewirkt; umgekehrt bei $p<q$. Wir haben festgesetzt, daß wir uns auf Werte von x beschränken wollen, für die

$$x\leqq\alpha\sqrt{npq},$$

wo α etwa mit 3 oder 4 angesetzt werden kann; daraus folgt, daß an der oberen Schranke von x

$$\frac{x}{np}=\alpha\sqrt{\frac{q}{np}} \qquad \frac{x}{nq}=\alpha\sqrt{\frac{p}{nq}},$$

daher

$$\frac{x}{nq}-\frac{x}{np}=\alpha\frac{p-q}{\sqrt{npq}}, \qquad \text{also} \qquad \frac{(p-q)x}{npq}=\frac{\alpha(p-q)}{\sqrt{npq}};$$

ist also $\frac{p-q}{\sqrt{npq}}$ sehr klein, so ist auch der Grad der Asymmetrie ein sehr schwacher; sieht man von ihm ab, so hat man

$$\frac{z}{z_0}=e^{-\frac{x^2}{2npq}}.$$

Die Asymmetrie fällt überhaupt fort bei $p=q=\frac{1}{2}$; für diesen Fall hat man also

$$\frac{}{}=e$$

Man hat hiernach als Grenze der binomialen Verteilung bei $p \neq q$ und einem sehr kleinen Wert von $\left|\frac{p-q}{\sqrt{npq}}\right|$:

$$z = z_0 e^{-\frac{x^2}{2npq}} \tag{4}$$

bei $p = q = \frac{1}{2}$, insbesondere

$$z = z_0 e^{-\frac{2x^2}{n}}. \tag{5}$$

Die Formeln tragen noch die Spuren ihrer Entstehung aus der Binomialreihe in den Größen p, q, n; man kann sie verwischen, wenn man beachtet, daß npq das Quadrat der mittleren Abweichung μ ist, das sich für $p = q = \frac{1}{2}$ auf $\frac{n}{4}$ reduziert; mithin erhält man für sehr kleine Werte von $\left|\frac{p-q}{\sqrt{npq}}\right|$ die Formel

$$z = z_0 e^{-\frac{x^2}{2\mu^2}}. \tag{6}$$

Die durch diese Gleichung dargestellte Kurve wird als normale Häufigkeitskurve, wegen ihres Auftretens in der Fehlertheorie, wo sie gleichfalls die Rolle einer Häufigkeitskurve spielt, auch als Fehlergesetzkurve oder kurz Fehlerkurve bezeichnet.

Ihre allgemeine Gestalt ist gekennzeichnet durch die Symmetrie bezüglich der Ordinatenachse und durch die asymptotische Annäherung an die Abszissenachse, die jedoch keine praktische Bedeutung hat, da nur solche Werte von x in Betracht kommen, die über das Drei- bis Vierfache μ nicht hinausgehen; endlich durch das Vorhandensein zweier Wendepunkte mit den Abszissen $\pm \mu$.

Die einzige Größe, von der die spezielle Gestalt abhängt, ist μ. Man kann sich jedoch auch von diesem Parameter freimachen, wenn man μ zur Maßeinheit für die Abszissen wählt und

$$\frac{x}{\mu} = \xi \tag{7}$$

setzt; alsdann wird für alle Verteilungen eine und dieselbe Häufigkeitskurve benützt werden können, die Kurve

$$z = z_0 e^{-\frac{\xi^2}{2}}. \tag{8}$$

Für z_0, die maximale Ordinate, haben wir in Art. 113 (4) den Näherungswert $\frac{1}{\sqrt{2\pi npq}}$ erhalten, wofür auch $\frac{1}{\mu\sqrt{2\pi}}$ geschrieben werden kann. Mithin hat man für die relative Verteilung der Abweichungen vom wahrscheinlichsten Werte np

$$z = \frac{1}{\mu\sqrt{2\pi}} e^{-\frac{x^2}{2\mu^2}} \tag{9a}$$

und für ihre absolute Verteilung

$$z = \frac{N}{\mu\sqrt{2\pi}} e^{-\frac{x^2}{2\mu^2}}. \tag{9b}$$

Man kann zu diesem letzten Ausdruck auch durch folgende Erwägung kommen. Die von der Verteilungskurve und der Abszissenachse begrenzte Fläche stellt den Umfang des Kollektivs dar; infolgedessen muß

$$\int_{-\infty}^{\infty} z\,dx = N$$

sein. Ersetzt man z durch den Ausdruck (6) und wendet hierauf die Substitution (7) an, so wird

$$z_0 \mu \int_{-\infty}^{\infty} e^{-\frac{\xi^2}{2}} d\xi = N$$

und mit der weitern Substitution $\frac{\xi}{\sqrt{2}} = t$

$$z_0 \mu \sqrt{2} \int_{-\infty}^{\infty} e^{-t^2} dt = N;$$

da nun das Integral den Wert $\sqrt{\pi}$ hat, so erhält man

$$z_0 = \frac{N}{\mu\sqrt{2\pi}}$$

in Übereinstimmung mit (10).

118. Die Auswertung der Ordinaten aller Verteilungskurven ist demnach auf die Werte der parameterfreien Funktion

$$\zeta = e^{-\frac{\xi^2}{2}}$$

zurückführbar; darum genügt es, eine Tabelle dieser Funktion herzustellen.[1] Eine solche, von 0,1 zu 0,1, in der letzten Einheit von 0,2 zu 0,2 fortschreitend, ist nachstehend mitgeteilt; neben den Werten von ζ sind auch deren Logarithmen angegeben.

[1]) Da es die Flächenverhältnisse an der Häufigkeitskurve sind, die hauptsächlich in Betracht kommen, so kann man auch vom Flächendifferential

$$z\,dx = \frac{1}{\mu\sqrt{2\pi}} e^{-\frac{x^2}{2\mu^2}} dx$$

ausgehen und dieses durch die Substitution $\frac{x}{\mu\sqrt{2}} = v$ parameterfrei machen; es geht nämlich über in

$$\frac{1}{\sqrt{\pi}} e^{-v^2} dv$$

und man kann nun $\frac{1}{\sqrt{\pi}} e^{-v^2}$ als Fehlerfunktion nehmen. Eine Tabelle dieser Funktion, nach Hundertsteln fortschreitend, vierstellig, von $v = 0{,}00$ bis $v = 3{,}06$ reichend und mit den ersten Differenzen versehen, findet sich in einer Abhandlung R. Wolfs in der Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich, 27. Jahrg., 1882, S. 258 und 259.

Tab. 66. Tafel der Funktion $\zeta = e^{-\xi^2}$

ξ	ζ	lg ζ	ξ	ζ	lg ζ
0,0	1,000 00	0	2,3	0,071 01	0,851 30 — 2
0,1	0,995 02	0,997 83 — 1	2,4	0,056 14	0,749 23 — 2
0,2	0,980 20	0,991 31 — 1	2,5	0,043 94	0,642 84 — 2
0,3	0,956 00	0,980 46 — 1	2,6	0,034 05	0,532 09 — 2
0,4	0,923 12	0,965 26 — 1	2,7	0,026 12	0,417 01 — 2
0,5	0,882 50	0,945 71 — 1	2,8	0,019 84	0,297 57 — 2
0,6	0,835 27	0,921 83 — 1	2,9	0,014 92	0,173 81 — 2
0,7	0,782 70	0,893 60 — 1	3,0	0,011 11	0,045 67 — 2
0,8	0,726 15	0,861 03 — 1	3,1	0,008 19	0,913 24 — 3
0,9	0,666 97	0,824 11 — 1	3,2	0,005 98	0,776 41 — 3
1,0	0,606 53	0,782 85 — 1	3,3	0,004 32	0,635 29 — 3
1,1	0,546 05	0,737 25 — 1	3,4	0,003 09	0,489 78 — 3
1,2	0,486 75	0,687 31 — 1	3,5	0,002 19	0,339 97 — 3
1,3	0,429 56	0,633 02 — 1	3,6	0,001 53	0,185 77 — 3
1,4	0,375 31	0,574 39 — 1	3,7	0,001 06	0,027 28 — 3
1,5	0,324 65	0,511 42 — 1	3,8	0,000 73	0,864 39 — 4
1,6	0,278 04	0,444 10 — 1	3,9	0,000 50	0,697 22 — 4
1,7	0,235 75	0,372 45 — 1	4,0	0,000 34	0,525 64 — 4
1,8	0,197 90	0,296 44 — 1	4,2	0,000 15	0,169 52 — 4
1,9	0,164 48	0,216 11 — 1	4,4	0,000 06	0,796 03 — 5
2,0	0,135 34	0,131 41 — 1	4,6	0 000 03	0,405 16 — 5
2,1	0,110 25	0,042 39 — 1	4,8	0,000 01	0,996 93 — 6
2,2	0,088 92	0,949 01 — 2	5,0	0,000 00	0.571 32 — 6

Zu beachten ist die anfänglich langsame, dann immer raschere Abnahme von ζ. Man müßte, wollte man die Kurve auch nur innerhalb des Bereichs $(-4,4)$ von ξ deutlich aufzeichnen, für die Ordinaten zum mindesten 1 m als Längeneinheit wählen; die kleinste Ordinate wäre dann etwa $\frac{1}{3}$ mm.

Die in der Fußnote [1]) auf S. 282 hergeleitete Fehlerfunktion $\varphi(v) = \frac{1}{\sqrt{\pi}} e^{-v^2}$, die man auch als Gaußsche Verteilung bezeichnet, begrenzt mit der v-Achse (Abszissenachse) eine Fläche von der Größe 1, da $\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \varphi(v)\, dv = 1$ ist. Das Moment zweiten Grades $M_2 = \mu^2$ (Quadrat der mittleren quadratischen Abweichung;

vgl. Art. 59) der Gaußschen Verteilung berechnet sich auf $\frac{1}{2}$. Das kann folgendermaßen bewiesen werden.

$$M_2' = \mu^2 = \int_{-\infty}^{\infty} v^2 \varphi(v)\, dv$$

Hieraus folgt

$$\mu^2 = \frac{1}{\sqrt{\pi}} \int^{\infty} v^2 e^{-v^2} dv.$$

Mittels der Methode der partiellen Integration erhält man, indem man zerlegt in $v \cdot v e^{-v^2} dv$,

$$\mu^2 = \frac{1}{\sqrt{\pi}} \left[-v \cdot \frac{1}{2} e^{-v^2} \right]_{-\infty}^{\infty} + \frac{1}{2\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-v^2} dv = \frac{1}{2\sqrt{\pi}} \sqrt{\pi} = \frac{1}{2}.$$

Das Moment vierten Grades stellt sich für die Gaußsche Verteilung auf $\frac{3}{4}$. Denn das Moment vierten Grades M_4' ist

$$M_4' = \int^{\infty} v^4 \varphi(v)\, dv.$$

Da $\varphi'(v) = -2v\varphi(v)$, also $\varphi(v) = -\frac{\varphi'(v)}{2v}$ ist, ergibt sich

$$M_4 = -\frac{1}{2} \int v^3 \varphi'(v)\, dv = -\frac{1}{2} \Big[v^3 \varphi(v) \Big]_{-\infty}^{\infty} + \frac{3}{2} \int_{-\infty}^{\infty} v^2 \varphi(v)\, dv = \frac{3}{2} \mu^2 = \frac{3}{2} \cdot \frac{1}{2} =$$

Durch Inbeziehungsetzen des Moments vierten Grades M_4' zum Moment zweiten Grades M_2' gelangt man für die Gaußsche Verteilung zu der Relation

$$\frac{M_4'}{M_2'^2} - 3 = \frac{\frac{3}{4}}{\frac{1}{4}} - 3 = 0.$$

Den Ausdruck $\frac{M_4'}{M_2'^2} - 3$ bezeichnet man als Steilheit (Exzeß) einer Verteilung.

$$\text{Exzeß} = \frac{M_4'}{M_2'^2} - 3 = \frac{M_4'}{\frac{1}{4}} - 3. \tag{10}$$

Für die Gaußsche Verteilung ist der Exzeß gleich 0.

Zur Bestimmung der Steilheit einer Verteilungskurve geht man in der Weise vor, daß man zunächst den Abszissenmaßstab so abändert, daß sich das Quadrat der mittleren quadratischen Abweichung ($\mu^2 = M_2'$) auf $\frac{1}{2}$ berechnet. Unter Zugrundelegung dieses neuen Maßstabes ermittelt man nach Formel (10) den Exzeß.

Dieser ist positiv, wenn die Verteilungskurve steiler verläuft als die entsprechende Gaußsche Verteilungskurve, im anderen Fall negativ.[1])

Zur Illustrierung wollen wir für zwei einfache diskontinuierliche oder arithmetische Verteilungen (vorstehende Betrachtungen bezogen sich auf kontinuierliche oder geometrische Verteilungen) den Exzeß bestimmen.

1. Beispiel. Wir markieren auf der Abszissenachse v vom Koordinatenanfangspunkt ausgehend nach beiden Seiten die Wechselpunkte (Klassengrenzen) $\frac{1}{2}$, $-\frac{1}{2}$, $1\frac{1}{2}$, $-1\frac{1}{2}$. Den Klassenmitten $+1$ und -1 ordnen wir die gleiche Häufigkeit 0,1 und dem Koordinatenanfang die Häufigkeit 0,8 zu. In unserem Beispiel fällt also das arithmetische Mittel in den Koordinatenanfang, und die numerischen Werte der Klassenmitten kennzeichnen zugleich die Abweichungen derselben vom arithmetischen Mittel. Es durchläuft v die Werte $+1$, 0 und -1 und $\varphi(v)$ die Werte 0,1, 0,8 und 0,1. Für μ^2 ergibt sich also

$$M_2' = \mu^2 = 1\ \ 0{,}1 + 0 \cdot 0{,}8 + 1 \cdot 0{,}1 = 0{,}2.$$

Um den Wert für das Quadrat der mittleren quadratischen Abweichung auf 0,5 zu bringen, hat man eine neue Abszisseneinheit einzuführen, die gleich ist $\sqrt{\frac{2}{5}}$ der alten Abszisseneinheit. In dieser neuen Abszisseneinheit stellen sich die Quadrate der Abweichungen der Klassenmitten auf 2,5, 0 und 2,5. Somit berechnet sich

$$M_2' = \mu^2 = 2{,}5 \cdot 0{,}1 + 0 \cdot 0{,}8 + 2{,}5 \cdot 0{,}1 = 0{,}5.$$

Das Moment vierten Grades M_4' beträgt also

$$M_4' = 2{,}5^2 \cdot 0{,}1 + 0 \cdot 0{,}8 + 2{,}5^2 \cdot 0{,}1 = 1{,}25.$$

Für den Exzeß unserer Verteilung erhalten wir

$$\text{Exzeß} = \frac{1{,}25}{0{,}5^2} - 3 = 2.$$

In unserem Beispiel ist also der Exzeß positiv, d. h. die Verteilungskurve verläuft steiler als die entsprechende Gaußsche Verteilungskurve.

2. Beispiel. Wir wählen dieselbe Klasseneinteilung wie im ersten Beispiel, ordnen aber den Klassenmitten $+1$ und -1 die gleiche Häufigkeit 0,3 und dem Koordinatenanfang die Häufigkeit 0,4 zu. Hiernach berechnet sich

$$M_2' = \mu^2 = 1 \cdot 0{,}3 + 0 \cdot 0{,}4 + 1 \cdot 0{,}3 = 0{,}6.$$

[1]) Vgl. R. v. Mises, Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik. Leipzig und Wien 1931, S. 242, und O. Anderson, Einführung in die mathematische Statistik. Wien 1935, S. 161 u. f.

Der neue Maßstab ist also $\sqrt{\frac{6}{5}}$ des alten. Im neuen Maßstab ausgedrückt, ergibt sich

$$M_2' = \mu^2 = \frac{5}{6} \cdot 0{,}3 + 0 \cdot 0{,}4 + \frac{5}{6} \cdot 0{,}3 = 0{,}5;$$

$$M_4' = \frac{25}{36} \cdot 0{,}3 + 0 \cdot 0{,}4 + \frac{25}{36} \cdot 0{,}3 = \frac{15}{36} = \frac{5}{12};$$

$$\text{Exzeß} = \frac{\frac{5}{12}}{\frac{1}{4}} - 3 = -\frac{4}{3}.$$

Der Exzeß ist negativ, die Verteilungskurve verläuft also weniger steil als die entsprechende Gaußsche Verteilungskurve.

119. Der in Art. 115 beschriebene Apparat zur mechanischen Herstellung der binomialen Verteilung würde bei großer Zahl der Keilreihen eine Annäherung an die Normalkurve ergeben. Doch sind der Vermehrung der Reihen konstruktive Grenzen gezogen.

Es verdient nun angemerkt zu werden, daß R. Wolf vor Erfindung des Galtonschen Apparates zu einer ähnlichen Erzeugung einer der Normalkurve angenäherten Linie gelangt ist.[1])

In ein Brettchen, das mit einer Zentimeterteilung versehen war, wurden zwei vertikale parallele Glastafeln im Abstand von 13.5 mm eingefügt; die Tafeln waren 38 cm lang und 9 cm hoch. Der so geschaffene Zwischenraum erhielt auf der einen Seite einen Abschluß durch eine Glaslamelle. Oben war ein trichterförmiger Aufsatz mit Ausflußöffnung verschiebbar angebracht, so daß er auf jeden Zentimeter der Teilung des Bodenbrettchens eingestellt werden konnte. Durch den Trichter wurden abgewogene Mengen eines sehr feinkörnigen trockenen Sandes eingeführt, die Begrenzungskurve bildete den Gegenstand der Messung und des Studiums. Es zeigte sich, daß sie bei Aufschüttungen, die über eine gewisse mäßige Höhe nicht hinausgingen, einer Normalkurve sehr nahe kam, sofern die seitliche Abschlußlamelle keine Stauwirkung üben konnte. Wurde aber der Fülltrichter der Seitenwand näher und näher gebracht, so gestaltete sich die Grenzkurve immer mehr asymmetrisch und schließlich selbst einseitig.

120. Bei der Ableitung der Exponentialfunktion aus der Binomialreihe ist n als eine sehr große Zahl vorausgesetzt worden und es ist zu erwarten, daß bei Zutreffen dieser Voraussetzung zwischen den Gliedern der Binomialreihe und den zugeordneten Funktionswerten nur sehr kleine Differenzen bestehen werden. Die praktische Durchrechnung zeigt aber, daß schon bei mäßig großem n eine befriedigende Übereinstimmung stattfindet. Das erhöht noch die Überlegenheit der Exponentialfunktion über die Binomialreihe.

Wir führen hier zum Nachweise des eben Gesagten zwei Beispiele vor; das eine betrifft eine symmetrische, das andere eine asymmetrische Binomialreihe; in letzterem Falle zeigt sich, wie die Exponentialfunktion die (schwache) Asymmetrie ausgleicht.

[1]) Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich, 27. Jahrg., 1882, S. 252 u. f.

1) Zur Binomialentwicklung $10\,000\left(\frac{1}{2}+\frac{1}{2}\right)^{32}$ gehört die mittlere Abweichung $\mu=\sqrt{32\cdot\frac{1}{2}\cdot\frac{1}{2}}=2{,}8284\ldots$; mit Hilfe derselben berechnet sich

$$z_0=\frac{10\,000}{\sqrt{16\,\pi}}=1410;$$

die Abweichungen

$$x= \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10,$$

in μ als Einheit ausgedrückt, ergeben

$$\xi= \quad 0{,}35 \quad 0{,}71 \quad 1{,}06 \quad 1{,}41 \quad 1{,}77 \quad 2{,}12 \quad 2{,}47 \quad 2{,}83 \quad 3{,}18 \quad 3{,}54;$$

mit Hilfe dieser Werte können auf Grund der Tab. 66 (Art. 118) mittels Interpolierens und des z_0 die Werte der Funktion z berechnet werden, die den Gliedern der Binomialentwicklung entsprechen. Dies führt zu nachstehender Zusammenstellung.

$x=\pm$	Binomialglied	Exponential-funktion	Unterschied in % des Binomialgliedes
0	1399	1410	0,79
1	1317	1325	0,61
2	1098	1096	0,18
3	809	804	0,62
4	526	522	0,76
5	300	295	1,67
6	150	149	0,67
7	65	67	3,08
8	24	26	8,33
9	8	9	12,50
10	2	3	50,00
11	—	—	—
Summe	9997	10002	

2) Die Entwicklung von $10\,000\left(\frac{2}{3}+\frac{1}{3}\right)^{42}$ liegt bereits ausgerechnet vor (Art. 112); jetzt handelt es sich um die zu ihr gehörige Exponentialfunktion. Dazu braucht man die mittlere Abweichung $\mu=\sqrt{42\cdot\frac{2}{3}\cdot\frac{1}{3}}=3{,}055$; mit dieser berechnet sich die Maximalordinate

$$z_0=10\,000\sqrt{\frac{3}{56\,\pi}}=1306;$$

die Abweichungen $x = 1, 2, 3, 4, \ldots$, in μ als Einheit ausgedrückt, geben $\xi = 0{,}33,\ 0{,}65,\ 0{,}98,\ 1{,}31, \ldots$; auf dieser Grundlage kann

$$z = z_0 e^{-\frac{\xi^2}{2}}$$

mit Hilfe unserer Tab. 66 mittels Interpolierens berechnet werden. Das Resultat ist folgendes:

x	Binomialglied	Exponentialfunktion	Unterschied in % des Binomialgliedes
+ 11	1	2	
+ 10	3	6	
+ 9	11	17	54,5
+ 8	33	42	27,3
+ 7	85	95	11,76
+ 6	185	192	3,78
+ 5	350	341	2,57
+ 4	578	554	4,15
+ 3	840	808	3,81
+ 2	1085	1057	2,58
+ 1	1252	1236	1,28
0	1297	1306	0,69
— 1	1211	1236	2,06
— 2	1022	1057	3,42
— 3	782	808	3,32
— 4	543	554	2,03
— 5	343	341	0,58
— 6	197	192	2,54
— 7	103	95	7,77
— 8	49	42	14,29
— 9	21	17	19,05
— 10	8	6	25,0
— 11	3	2	33,3
— 12	1	—	—
Summe	1003	10006	

121. Die Anwendbarkeit der normalen Häufigkeitskurve beschränkt sich nicht auf Verteilungen, die ihrer Natur nach unter die Binomialformel fallen, wie es Versuche mit Münzen, Würfeln, Ziehungen aus Urnen mit verschieden gefärbten Kugeln u. ä. sind, kurz auf Gesamtheiten, bei welchen es sich um die Häufigkeit des Auftretens und Fehlens eines bestimmt umschriebenen Merkmals handelt. Sie erstreckt sich auch auf Kollektive, deren Glieder mit einem variablen, in erster Linie mit einem quantitativen Merkmal behaftet sind, das durch eine stetige

Variable ausdrückbar ist. Bei solchen Materien ist eine Beziehung zum Binomialtheorem nicht unmittelbar ersichtlich und könnte nur auf künstlichem Wege, durch Vermittlung einer Hypothese, hergestellt werden.

Dies ist auf einem Gebiete geschehen, von welchem die normale Häufigkeitskurve ihren Ursprung nahm: in der Fehlertheorie. Der zufällige Fehler ist als eine stetige Variable anzusehen, und die Verteilung der Werte, die er in einer Reihe von Beobachtungen annehmen kann, folgt nach vielfachen Erfahrungen dem normalen Häufigkeitsgesetz, wenn nur gewisse Voraussetzungen erfüllt sind.

Um diesen Sachverhalt mathematisch zu begründen, hat man verschiedene Hypothesen über die Entstehung der Beobachtungsfehler gemacht, deren eine am meisten Anerkennung gefunden hat, weil sie sich offenkundig der Wirklichkeit anpaßt; dieser Hypothese zufolge ist der Beobachtungsfehler nicht eine einfache, sondern eine komplexe Größe, zusammengesetzt aus den Wirkungen zahlreicher Fehlerquellen, deren jede für sich nur einen sehr kleinen Einfluß auszuüben vermag; der Beobachtungsfehler erscheint darnach als algebraische Summe einer Anzahl von Variablen (Elementarfehlern), deren jede positive und negative Werte aus einem sehr engen Intervall annehmen kann. Die analytische Verarbeitung dieser Hypothese führt zur normalen Verteilungsfunktion als Häufigkeitsgesetz der resultierenden Beobachtungsfehler. Die Hypothese ist immer allgemeiner gefaßt worden: so ist man von der ursprünglichen Annahme, die Elementarfehler seien gleich groß, negativ oder positiv mit gleicher Wahrscheinlichkeit, abgegangen und hat bei ihnen beliebige Wirkungsgesetze vorausgesetzt; man hat weiter erkannt, daß die Voraussetzung völliger gegenseitiger Unabhängigkeit der Fehlerquellen nicht notwendig ist.

Dieser hohe Grad von Allgemeinheit erleichtert es, die eben entwickelten Vorstellungen auf andere Kollektive zu übertragen, an denen Verteilungen nach der normalen Häufigkeitskurve beobachtet worden sind. Es gehören hierher vor allem andern anthropologische Gegenstände, Maße und Gewichte am menschlichen Körper. Nehmen wir beispielsweise als eine der zu allererst untersuchten Größen die Körperhöhe. Ihre Komplexität ist nicht zu leugnen; sie erscheint als Summe einer großen Zahl von Einzeldimensionen, von denen jede für sich zur Ausbildung kommt, wenn auch nicht in völliger Unabhängigkeit von den andern, so doch auch nicht in völliger Korrelation zu ihnen; daneben wirken andere Umstände mit, so Krümmungsverhältnisse, Gewohnheiten in der Körperhaltung u. a. m. Auch im Tier- und Pflanzenreich sind vielfach normale Verteilungen festgestellt worden und lassen sich auch da durch analoge Erwägungen erklären.

Demgegenüber gibt es Gebiete, wo Verteilungen nach dem normalen Häufigkeitsgesetz nur sehr selten oder gar nicht anzutreffen sind. Dies gilt beispielsweise von wirtschaftsstatistischen Erfahrungsreihen, z. B. Reihen von Produktionszahlen, von Warenpreisen, von Geldsätzen. Wohl wirken auch hier viele Ursachen zusammen, aber die Voraussetzung, daß sie sich in nahe gleichen Grenzen halten, daß sie bald in dem einen, bald in dem andern Sinne wirken, trifft nicht zu: oft ist eine einzige Ursache derart überwiegend, daß sie der ganzen Erscheinung ihr Gepräge verleiht und die andern zurückdrängt.

Dasselbe läßt sich von manchen Erfahrungsreihen im land- und forstwirtschaftlichen Betrieb sagen. So hängen die Bodenerträge auf verschiedenen Parzellen von zahlreichen Wachstumsfaktoren ab, von denen ein einzelner so durchschlagend werden kann, daß alle andern ihm gegenüber fast völlig zurücktreten.

Diese Überlegungen sind zu dem Zwecke angestellt worden, um das normale Häufigkeitsgesetz aus seiner Verbindung mit der binomialen Verteilung zu lösen, in die es hier durch die Art seiner mathematischen Abteilung gekommen ist, und begreiflich zu machen, daß es auch in Fällen auftreten kann, die zu dem Binomialtheorem keine ersichtliche Beziehung aufweisen.

122. Zwischen der Theorie der Kollektive und der Fehlertheorie bestehen mancherlei Beziehungen, die zum Teil ihren Grund darin haben, daß die normale Häufigkeitskurve in der Fehlertheorie unter dem Namen Fehlerwahrscheinlichkeitskurve oder Gaußsches Fehlergesetz eine so maßgebende Rolle spielt. Aber aus einer zu weitgehenden Übertragung dessen, was in der Fehlertheorie seit langem ausgebildet war, auf Kollektive der verschiedensten Art haben sich allerhand Mißverständnisse, Mißdeutungen ergeben, auf die hier mit einigen Worten hingewiesen werden soll.

Man hat anfänglich ohne nähere Prüfung als selbstverständlich angenommen, daß eine Kollektivreihe, entstanden durch Messung eines quantitativen Merkmals an einer Reihe gleichartiger Gegenstände, in allen Belangen vergleichbar sei einer Beobachtungsreihe, entstanden durch wiederholte Messung einer und derselben Größe unter völlig gleichen Umständen. Daß Quetelet bei seinen ersten derartigen Untersuehungen auf eine Materie stieß — Brustumfänge und Körperhöhen — bei welcher dies in hohem Grade zutrifft, hat zur Festigung dieser Meinung beigetragen.

Nun aber liegen die Dinge in den beiden Fällen ganz verschieden.

Das eine Mal handelt es sich um eine mehr oder minder scharf definierte Größe, deren Wert man bestimmen will und an der man die Messung wiederholt, um die Genauigkeit der Bestimmung zu erhöhen und ein Maß für sie zu erforschen. Die unvermeidlichen zufälligen Fehler, die die Erkenntnis des wahren Wertes verhindern, sind maßgebend für die erzielte Genauigkeit des Resultates, für dessen Annahme man sich entschieden hat.

Im andern Falle mißt man die verschiedenen Variationen eines und desselben Merkmals, mit denen es an den Gliedern eines Kollektivs auftritt. Die Verschiedenheit der Werte, die dabei erhalten werden und die dort als ein Mangel, als eine Unvollkommenheit unserer Messungen empfunden werden mußte, gibt hier Aufklärung darüber, welche Beständigkeit das Merkmal innerhalb des Kollektivs aufweist.

Was also dort ein Mittel zur Erkenntnis der Genauigkeit ist, ist hier ein Mittel zur Beurteilung der Beständigkeit.

Auf diesen wesentlichen Unterschied ist meist nicht geachtet worden und darum haben sich die Bezeichnungen: Fehler, Genauigkeit aus der Fehlertheorie in die Lehre von den Kollektiven verpflanzt, wo sie keine Berechtigung haben.

Man kann allerdings von Fehlern und von Genauigkeit bei der Erhebung eines Kollektivs sprechen; denn diese setzt sich aus Beobachtungen (Messungen, Wägungen, Zählungen) zusammen, und bei diesen kommen Fehler vor, und es ist mit verschiedenen Graden von Genauigkeit zu rechnen. Wollte man diese erproben, so müßte dies gesondert geschehen nach den Methoden, welche in der Meßkunde zur Genauigkeitsbestimmung eines Meßwerkzeugs, eines Meßverfahrens angewendet werden und in der Hauptsache in deren wiederholter Anwendung auf ein und dieselbe Größe bestehen. Nebenbei sei bemerkt, daß zu den Beobachtungsfehlern auch der Umstand zu rechnen wäre, daß man bei der Klassen-

bildung Glieder verschiedener Maße in einer Klasse zusammenfaßt und als gleich behandelt. Der Einfluß, den dies z. B. auf die mittlere Abweichung hat, ist besonders untersucht worden und die Sheppardsche Korrektur bringt ihn in Rechnung (Art. 63). Doch darf nicht übersehen werden, daß auch diese Korrektur auf Annahmen beruht, von welchen man sich nur ein angenähertes Zutreffen versprechen darf. Aus diesem Vorgang allein entspringen schon Ungenauigkeiten, die die eigentlichen Beobachtungsfehler meist übertreffen dürften.

Aber das Interesse richtet sich nicht auf die von den Erhebungsfehlern stammenden Ungleichheiten, sondern auf die im Wesen des Kollektivs ruhenden Schwankungen. In der Regel, wenn auch vielleicht nicht immer mit Berechtigung, nimmt man an, daß jene Fehler gegenüber diesen Schwankungen derart zurücktreten, daß man von ihnen absehen kann. Wenn z. B. mit einem der hierzu konstruierten Meßwerkzeuge bestimmte Dimensionen an den Köpfen einer bestimmten Menschengruppe (einem Kollektiv) gemessen werden, so haftet jeder solchen Messung eine von dem Werkzeug und dessen Handhabung stammende Ungenauigkeit, besser Unsicherheit, an; da aber die Unterschiede an den verschiedenen Köpfen viel erheblicher sind als diese kleine Unsicherheit ausmachen kann, so wird meist von ihr abgesehen. Man kann sie aber, wie schon bemerkt, durch ein Vorverfahren abschätzen und dann in Rechnung bringen, um einen Ausdruck für die von den Meßfehlern befreite, also für die in der Natur des Kollektivs begründete Variabilität der Individuen in dem betreffenden Merkmal zu erhalten.

Käme es nur auf eine ungerechtfertigte Übertragung von Namen an, so könnte man sich darüber hinwegsetzen; denn das ist nichts Ungewöhnliches, daß man mit demselben Namen verschiedene Begriffe verbindet. Aber bedenklich ist es, wenn Formeln, Schlußfolgerungen auf Gegenstände angewendet werden, wo die Voraussetzungen dafür nicht erfüllt sind.

Es ist viel Arbeit und Mühe darauf verwendet worden, zu prüfen, ob die Ergebnisse landwirtschaftlicher Versuche, Vegetationsversuche wie Anbauversuche im freien Felde, sich der normalen Verteilung fügen. Der Erfolg solcher Prüfungen war verschieden, wurde aber zumeist als befriedigend erachtet und daraus die Anwendbarkeit der Fehlertheorie, der Methode der kleinsten Quadrate, auf das landwirtschaftliche Versuchswesen gefolgert.[1]) In der Folge hat H. Vater die prüfende Untersuchung auch auf forstliche Versuche ausgedehnt.[2]) Er zieht aus einer Zusammenstellung älterer landwirtschaftlicher Kollektivreihen (zum Teil sind es die unter [1]) besprochenen) den nicht ganz gerechtfertigten Schluß, es sei als er-

[1]) Eine kritische Darstellung einer Reihe solcher Bestrebungen hat E. Czuber in der „Zeitschrift für das landwirtschaftliche Versuchswesen in Österreich", Jahrg. 1918, S. 1 u. f. gegeben.

[2]) H. Vater, Die Ausgleichungsrechnung bei Bodenkulturversuchen, Mitteilungen aus der Sächsischen forstlichen Versuchsanstalt zu Tharandt, Bd. II, S. 1 u. f. Mit der Bezeichnung „Bodenkultur" soll sowohl Land- als auch Forstwirtschaft getroffen werden. Der Verfasser führt wohl auch Versuchsreihen aus der Landwirtschaft an, aber hauptsächlich sind es von ihm selbst angestellte forstliche Versuche, auf die er seine Untersuchungen gründet. Auf der einen Seite sind es Saatversuche zur Feststellung der Zulänglichkeit der Nährstoffe des Waldbodens für das Gedeihen von Kiefer und Fichte; erhoben wird das Trockengewicht der Saatpflanzen. Auf der anderen Seite betreffen die Versuche Jungwüchse verschiedenen Alters von Kiefer und Fichte; gemessen wird die Höhe der Bäumchen.

wiesen anzusehen, daß Wachstumabweichungen das Gaußsche Gesetz befolgen; die Reihen gehen von 24 bis auf 500 Glieder. Minder positiv spricht sich H. Vater bezüglich seiner forstlichen Versuche aus, wo sich deutliche Asymmetrien ausgesprochener Richtung bemerkbar machen.

Nun bleibt, selbst wenn die normale Verteilung wirklich vorhanden wäre, die Frage offen, bei welcher Anzahl von Versuchen man erwarten darf, daß sie zu den die betreffende Verteilung charakterisierenden Größen mit solcher Annäherung führen, um alle Folgerungen aus dem Fehlergesetz mit Vertrauen auf ihr Zutreffen ziehen zu können.

In dieser Beziehung sei erwähnt, daß man im landwirtschaftlichen Versuchswesen selbst ganz kurze Erfahrungsreihen, sogar bis zu zwei Gliedern herab, glaubte, zur Grundlage nehmen zu können.

Die Frage hat nicht nur theoretische, sondern auch große praktische Bedeutung, denn Versuche kosten Zeit und Geld.

Vater hat auf experimentellem Wege gesucht, zu einer Antwort zu kommen. Aus Wertreihen, die sich auf eine (genau) bekannte Größe beziehen und das Gaußsche Gesetz befolgen, hob er nach Zufall Gruppen verschiedener Gliederzahl heraus und prüfte daran, wie nahe sie die Beziehungen erfüllen, die aus dem Gaußschen Gesetz fließen, also z. B. das Verhältnis der wahren Abweichung zur mittleren, das Verhältnis der durchschnittlichen zur mittleren. Er zog aus den mühevollen Rechnungen die Vermutung, daß erst bei 50 bis 100 Erfahrungsergebnissen eine genügend starke Annäherung erhofft werden darf.

Darin liegt ein Hinweis darauf, daß es nicht gerechtfertigt ist, wenn auf Reihen, die aus Bodenkulturversuchen hervorgegangen sind, vorbehaltlos und ohne Rücksicht auf Umfang weitgehende Wahrscheinlichkeitsschlüsse gestützt werden, die nur Bestand haben, wo das Gaußsche Gesetz in aller Strenge gilt.

Von diesem Standpunkte aus sind auch verschiedene Rechnungsweisen zu beurteilen, die man angewendet hat. Man hat beispielsweise die durchschnittliche Abweichung einer Kollektivreihe bestimmt als Korrelat des durchschnittlichen Fehlers und hat daraus rechnerisch die wahrscheinliche Abweichung abgeleitet, und so wurde auch mit der mittleren Abweichung verfahren; ganz in der Art, als handelte es sich um den durchschnittlichen, mittleren und wahrscheinlichen Fehler im strengen Sinne der Fehlertheorie. Das hängt damit zusammen, daß man in der Landwirtschaft den wahrscheinlichen Fehler wegen seiner leichten Verständlichkeit bevorzugt, wie in manchen anderen Gebieten.

Die verschiedenen Streuungsmaße, deren sich die Theorie der Kollektive bedient, sind aber nicht auf gleiche Stufe zu stellen mit den Genauigkeitsmaßen der Fehlertheorie. Die Beziehungen, die unter diesen bei Geltung des Gaußschen Gesetzes stattfinden, sind auf jene nicht übertragbar. Die Streuungsmaße behalten ihre Bedeutung bei jeder Art von Verteilung, ihre Beziehungen zueinander wechseln aber mit der Verteilung. Ihre Bestimmung muß unabhängig erfolgen, wenn sie einen Aufschluß geben sollen über die Art der Verteilung. Es läßt sich von vornherein nicht sagen, ob man bezüglich zweier Kollektive zu demselben Urteil über die Größe der Streuung kommt, wenn man mittlere, durchschnittliche Abweichungen oder Quartile verwendet. Darum empfiehlt sich eine einheitliche Vereinbarung; die mittlere Abweichung hat gleich dem mittleren Fehler in der Fehlertheorie bisher den Vorzug erhalten.

Nach mancherlei Versuchen führen bei Verteilungen, die auch nur die Hauptzüge der normalen an sich tragen, alle Streuungsmaße meist zu demselben Ur-

teil (Art. 66). Bei Asymmetrie hätte eine einheitliche wahrscheinliche Abweichung keine Berechtigung, weil sich die beiden Seiten verschieden verhalten.

Die mittlere Abweichung bezieht sich ihrer Definition gemäß auf das arithmetische Mittel als Ausgangswert; sie ist die Quadratwurzel aus dem Durchschnitt der Quadrate der Abweichungen der Kollektivglieder von diesem. Man begegnet aber auch der Formel, die in der Fehlertheorie als Ausdruck des mittleren Fehlers einer Beobachtung bekannt ist und sich von der obigen Definition nur durch den Nenner unterscheidet, indem als solcher die um 1 verminderte Gliederzahl auftritt. Der Grund ist der, daß man hier nach der mittleren Abweichung vom wahren Werte sucht, den man sich in der gemessenen Größe tatsächlich verwirklicht denkt. Eine solche Verwirklichung fehlt beim Kollektiv, daher hat die Bezugnahme darauf zu entfallen. Man begegnet Darstellungen, die sich so lesen, als ob mit dem Nenner $n-1$ eine Verschärfung in der Bestimmung einträte, während die Beibehaltung von n eine bloße Näherung bedeuten würde.

Wenn im vorstehenden von der Methode der kleinsten Quadrate Gebrauch gemacht wurde, wie bei der Aufstellung der Regressionsgleichungen, so hat dies mit dem Fehlergesetz und mit Wahrscheinlichkeit nichts zu tun; vielmehr gilt dabei die Methode als ein Prinzip, nach welchem man die „möglichste Erfüllung“ von Gleichungen durch ein System beobachteter Werte bewirkt.

123. Zu unserem Gegenstande zurückkehrend, haben wir uns jetzt mit der Aufgabe zu befassen, zu einer vorliegenden Verteilung, die dem Anscheine nach die Anpassung an eine Normalkurve erwarten läßt, diese letztere zu bestimmen. Dazu ist vor allem ein hoher Grad von Symmetrie, insbesondere im Kernstück der Verteilung, notwendige Voraussetzung. Ist die Kurve gefunden, so wird es sich noch darum handeln, über den Grad der Anpassung eine Prüfung anzustellen.

Die dazu nötigen Vorkehrungen und Rechnungen werden sich am besten im Anschlusse an ein Beispiel darlegen lassen. Wir wählen dazu die Verteilung der Körperhöhen amerikanischer Rekruten (Art. 60, 1). Von dorther nehmen wir die folgenden Größen:

$$N = 25878$$
$$M = 66{,}7011''$$
$$\mu = 2{,}5524''.$$

Daraus berechnet sich die Maximalordinate der Normalkurve

$$z_0 = \frac{N}{\mu\sqrt{2\pi}} = 4045,$$

mithin lautet die Gleichung der Kurve selbst

$$z = 4045\, e^{-\frac{\xi^2}{2}};$$

dabei bedeutet ξ die in μ als Einheit ausgedrückte Abszisse; Ursprung ist der Punkt M, der hier zugleich den Zentralwert und den dichtesten Wert vorzustellen hat. Damit ist die Normalkurve der Form und Lage nach bestimmt. Ihre Auf-

zeichnung geschieht durch Ausrechnung einer genügend gro[illegible] Anzahl von Ordinaten, durch deren Endpunkte dann die Kurve zu ziehe[illegible]. Bei dieser Ausrechnung leistet die Tab. 66 (Art. 118) gute Dienste, die [illegible]iplikationen können überdies durch Benützung der Logarithmen umgangen [illegible]en. Hat man die Abszissen gewählt — man läßt sie zweckmäßig mit den Klassenmitten zusammenfallen — so ist ihre Reduktion auf μ als Einheit durchzuführen. Nachstehend sind die Werte von x, ξ und ζ zusammengestellt.

$x = \pm$	ξ	ζ
0	0	1
1	0,39179	0,92613
2	0,78358	0,73565
3	1,17536	0,50121
4	1,56715	0,29289
5	1,95894	0,14679
6	2,35073	0,06311
7	2,74252	0,02327
8	3,13430	0,00736
9	3,52609	0,00200
10	3,91788	0,00046

Hiernach stellt sich der wirklichen Verteilung die nebenstehende theoretische gegenüber.

Klassenmitte in Zoll	z beobachtet	z berechnet
51,5	1	—
52,5	1	—
53,5	2	—
54,5	1	—
55,5	3	—
56,5	7	2
57,5	6	8
58,5	10	30
59,5	15	94
60,5	50	255
61,5	526	594
62,5	1237	1185
63,5	1947	2027
64,5	3019	[illegible]76
65,5	3475	[illegible]
66,5	4054	4[illegible]
67,5	3631	37[illegible]
68,5	3133	2976
69,5	2075	2027
70,5	1485	1185
71,5	680	594
72,5	343	255
73,5	118	94
74,5	42	30
75,5	9	8
76,5	6	2
77,5	2	—
	25878	25879

Die Übereinstimmung kann nicht in allen Teilen als befriedigend bezeichnet werden; das war nicht anders zu erwarten, da die beobachtete Verteilung einzelne auffallende Unregelmäßigkeiten zeigt. Aber immerhin kann nicht in Abrede gestellt werden, daß sich der normale Verteilungstypus deutlich verrät.

Dies tritt klarer hervor, wenn man das wirkliche Häufigkeitspolygon mit der Normalkurve zusammen zur Darstellung bringt. Um Vergleichbarkeit zu erzielen, müssen für beide gleiche Maßstäbe gewählt werden.

Der eine Maßstab in der Abszissenachse betrifft die Körperhöhe. Als Einheit ist hier μ angenommen. Stellt man es durch die Länge 1 cm dar, so ist das Klassenintervall $\frac{1}{\text{''}}$, also 0,392 cm.

Der andere Maßstab, in der dazu senkrechten Lage, ist ein Häufigkeitsmaßstab, und es handelt sich darum, wie viele Individuen auf die gewählte Maßeinheit entfallen. Geht man von der größten Ordinate z_0 aus und gibt ihr die Länge von 6 cm, so ist alles bestimmt, und zwar durch die Forderung, daß die Fläche des Polygons gleich sein muß der Fläche der Kurve, beide darstellend die Gesamtzahl N der Individuen. Nun ist die Fläche der Kurve, wenn μ als Einheit genommen wird,

$$\int_{-\infty}^{\infty} e^{-\frac{\xi^2}{2}} d\xi = z_0 \sqrt{2} \int_{-\infty}^{\infty} e^{-t^2} dt = z_0 \sqrt{2\pi},$$

bei unserer Annahme also $6\sqrt{2\pi} = 15{,}04$; auf die Flächeneinheit entfallen $n \cdot 2{,}55$ Individuen, wenn der zweite Maßstab n Individuen pro Längeneinheit, d. i. pro Zentimeter bedeutet; mithin hat man zur Bestimmung von n den Ansatz:

$$\frac{25878}{n \cdot 2{,}55} = 15{,}04,$$

woraus $n = 674{,}7$. Es bedeutet also 1 cm der zweiten Skala 674,7 Individuen oder 1000 Individuen entsprechen 14,82 mm.

Man erhält schließlich die Ordinaten der Polygonecken in Zentimetern, indem man die Klassenhäufigkeiten durch 674,7 dividiert und die entsprechenden Ordinaten der Normalkurve, indem man die obigen ζ mit 6 multipliziert. Die Werte sind auf Seite 296 zusammengestellt und dann zur Konstruktion der Fig. 36 ver-

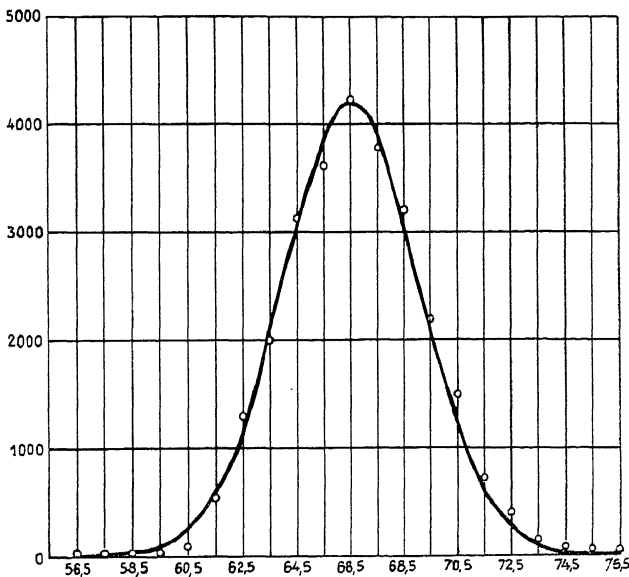



Fig. 36. Anpassung der Normalkurve an eine gegebene Verteilung.

wendet worden. Das Polygon selbst ist nicht eingezeichnet, nur seine Ecken sind durch geringelte Punkte sichtlich gemacht. Der Ursprung der Normalkurve ist gegenüber dem Nullpunkt des Polygons nach rechts verschoben, entsprechend dem Unterschied $M - 66{,}5 = 0{,}2011$, in unserer Einheit $= 0{,}079$ cm.

Klassenmitte in Zoll	Ordinaten in Zentimeter		Klassenmitte in Zoll	Ordinaten in Zentimeter	
	Polygon	Normalkurve		Polygon	Normalkurve
56,5	0,010	0,003	67,5	5,38	5,56
57,5	0,009	0,012	68,5	4,64	4,41
58,5	0,015	0,044	69,5	3,08	3,01
59,5	0,022	0,14	70,5	2,20	1,76
60,5	0,074	0,38	71,5	1,01	0,88
61,5	0,78	0,88	72,5	0,51	0,38
62,5	1,83	1,76	73,5	0,17	0,14
63,5	2,89	3,01	74,5	0,062	0,044
64,5	4,47	4,41	75,5	0,013	0,012
65,5	5,15	5,56	76,5	0,009	0,003
66,5	6,01	6,00	77,5	0,003	.
				38,3	38,4

124. Die Prüfung des Anschlusses einer Normalkurve an die ihr zugrunde liegende Verteilung ins einzelne geschieht in der Weise, daß man die gerechnete Häufigkeit für bestimmte Intervalle vergleicht mit der wirklich beobachteten. Die Intervallgrenzen läßt man, um Interpolationen zu vermeiden, tunlichst mit Klassengrenzen zusammenfallen. Das setzt die Quadratur der Normalkurve über beliebigen Teilen der Abszissenachse voraus. Alle auftretenden Fragen lassen sich erledigen, wenn man zu einer genügend dichten Reihe von Ordinaten jedesmal eine der beiden Flächen kennt, in welche sie die Gesamtfläche teilen; es soll die größere von beiden gewählt und mit $\Psi(\xi)$ bezeichnet werden, wobei ξ die in μ ausgedrückte Abszisse bedeutet.

Gewöhnlich findet man in den Schriften über Wahrscheinlichkeitsrechnung und Fehlertheorie Tabellen der Funktion

$$\Phi(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-u^2} du, \tag{11}$$

welche die Fläche über dem Intervall $(-x, x)$ kennzeichnet, wenn $t = \frac{x}{\mu\sqrt{2}} = \frac{\xi}{\sqrt{2}}$ ist: mit dieser hängt die Funktion Ψ in der Weise zusammen, daß

$$\Psi(t) = \frac{1 + \Phi(t)}{2} \tag{12a}$$

ist. Dieser Formel gemäß ist die folgende Tabelle aus einer Tafel von $\Phi(t)$ abgeleitet worden in einem Umfange, der für viele praktische Zwecke ausreichen dürfte.

Tab. 67. Tafel der größeren von den durch ξ bestimmten Flächen.

ξ	$\Psi\left(\frac{\xi}{\sqrt{2}}\right)$	ξ	$\Psi\left(\frac{\xi}{\sqrt{2}}\right)$	ξ	$\Psi\left(\frac{\xi}{\sqrt{2}}\right)$	ξ	$\Psi\left(\frac{\xi}{\sqrt{2}}\right)$
0,00	0,500 00	1,00	0,841 34	2,00	0,977 25	3,00	0,998 65
0,05	0,519 97	1,05	0,853 15	2,05	0,979 82	3,05	0,998 86
0,10	0,539 83	1,10	0,864 33	2,10	0,982 14	3,10	0,999 03
0,15	0,559 63	1,15	0,875 11	2,15	0,984 22	3,15	0,999 19
0,20	0,579 26	1,20	0,884 93	2,20	0,986 10	3,20	0,999 31
0,25	0,598 72	1,25	0,894 35	2,25	0,987 78	3,25	0,999 42
0,30	0,617 91	1,30	0,903 20	2,30	0,989 28	3,30	0,999 52
0,35	0,636 83	1,35	0,911 49	2,35	0,990 61	3,35	0,999 59
0,40	0,655 42	1,40	0,919 24	2,40	0,991 80	3,40	0,999 66
0,45	0,673 65	1,45	0,926 46	2,45	0,992 85	3,45	0,999 72
0,50	0,691 46	1,50	0,933 19	2,50	0,993 79	3,50	0,999 77
0,55	0,708 83	1,55	0,939 42	2,55	0,994 61	3,55	0,999 81
0,60	0,725 75	1,60	0,945 20	2,60	0,995 34	3,60	0,999 84
0,65	0,742 14	1,65	0,950 52	2,65	0,995 98	3,65	0,999 87
0,70	0,758 04	1,70	0,955 43	2,70	0,996 53	3,70	0,999 89
0,75	0,773 36	1,75	0,959 93	2,75	0,997 02	3,75	0,999 91
0,80	0,788 14	1,80	0,964 07	2,80	0,997 44	3,80	0,999 93
0,85	0,802 32	1,85	0,967 84	2,85	0,997 81	3,85	0,999 94
0,90	0,815 94	1,90	0,971 28	2,90	0,998 13	3,90	0,999 95
0,95	0,828 95	1,95	0,974 41	2,95	0,998 41	3,95	0,999 96
						4,00	0,999 97

Umgekehrt können aus dieser Tabelle die Werte von $\Phi(t)$ abgeleitet werden nach der Formel

$$\Phi(t) = 2\Psi(t) - 1. \tag{12b}$$

So ergeben sich für

$$\xi = \quad 1 \qquad 2 \qquad 3 \qquad 4$$

die Werte

$$\Phi\left(\frac{\xi}{\sqrt{2}}\right) = \quad 0{,}68268 \qquad 0{,}95450 \qquad 0{,}99730 \qquad 0{,}99994,$$

welche die Häufigkeit in den Intervallen

$$(-\mu, \mu) \qquad (-2\mu, 2\mu) \qquad (-3\mu, 3\mu) \qquad (-4\mu, 4\mu)\ ^{1})$$

angeben.

[1]) C. Boehm verwendet diese Ergebnisse für die Behandlung der Frage der Schwankungsreserve in der Lebensversicherung. Er setzt die Schwankungsreserve gleich dem Vierfachen der Streuung. Vgl. C. Boehm, Das Problem der „richtigen" Sterbetafel. Archiv für mathematische Wirtschafts- und Sozialforschung 1936, Bd. II, Heft 3, S. 173.

In einer 1000gliedrigen Reihe ist also zu erwarten, daß 997 Fälle innerhalb der Grenzen -3μ und $+3\mu$, also nur 3 ‰ aller Fälle außerhalb dieser Grenzen fallen werden. Darauf gründet sich der häufig gebrauchte Schluß, daß Abweichungen, die dem Betrage nach die dreifache mittlere Abweichung erheblich überschreiten, als zufällige Störungen nicht zu erwarten sind, und die Regel, daß in einem Kollektiv die Hauptmasse der Abweichungen — bis auf einen geringen Promillesatz — innerhalb eines Intervalls von der sechsfachen Ausdehnung der mittleren Abweichung zu suchen ist. Man darf bei den Anwendungen dieser Regeln nicht übersehen, daß sie aus den Gesetzen der Normalkurve stammen und daher nur für dieser nahekommende Verteilungen Bestätigung erwarten lassen.

Die vorstehenden Betrachtungen spielen auch in der Technik eine wichtige Rolle. Es sei in diesem Zusammenhang auf die Untersuchungen von Becker[1]), Plaut[1]) und Runge[1]) hingewiesen. Plaut (Fabrikationskontrolle, S. 33) beschäftigt sich u. a. mit den Versagern, die zuweilen bei Sprenggeschossen auftreten, und behandelt die Frage, wie viele Geschosse abgeliefert werden müssen, um den Prozentsatz der Versager in der Lieferung mit hinreichender Sicherheit und Genauigkeit festzustellen. Weiter erörtert Plaut auch die Frage, ob eine neue Verpackung hinreichend gegen Bruch auf dem Transport schützt. Hierbei nimmt Plaut an, daß der Schutz hinreichend ist, wenn nicht mehr als 1 % der Stückzahl auf dem Transport beschädigt werden. Plaut setzt an, daß sich in einer Sendung von 5000 Stück 30 beschädigte Exemplare vorfinden. Den Bereich der Zufallsschwankungen setzt man mit dem $3^1/_2$fachen Betrag der theoretischen mittleren Abweichung an. Der Schwankungsbereich berechnet sich auf

$$0{,}6 \pm 3{,}5\sqrt{\frac{0{,}6\,(100-0{,}6)}{5000+3}} = 0{,}6 \pm 0{,}39.$$

Die Verpackung genügt noch der gestellten Bedingung.

125. Bei der zuletzt untersuchten Verteilung der Körperhöhen amerikanischer Rekruten fällt die Größenklasse 68 bis 69, noch mehr aber die Klasse 60 bis 61 Zoll durch eine verhältnismäßig große Abweichung der beobachteten Häufigkeit von der berechneten auf, wie auch Fig. 36 erkennen läßt. Es soll geprüft werden, ob diese Abweichungen noch als zufällige gelten können.

Die Enden der Klasse 68 bis 69 sind von $M = 66{,}7011$ um 1,2989, bzw. 2,2989 entfernt; in der Einheit $\mu = 2{,}5524$ ausgedrückt, sind diese Abweichungen 0,509, bzw. 0,901.

Die zu diesen Zahlen gehörigen Tabellenwerte sind, bei Beschränkung auf lineare Interpolation, 0,69459, bzw. 0,81620; ihre Differenz 0,12161 gibt die relative und

$$0{,}12161 \cdot 25878 = 3147$$

die absolute Häufigkeit in der betreffenden Klasse, während die beobachtete 3133 war; das gibt die Differenz 14.

[1]) R. Becker, H. Plaut und J. Runge, Anwendungen der mathematischen Statistik auf Probleme der Massenfabrikation, Berlin 1930. Vgl. auch H. Plaut, Fabrikationskontrolle auf Grund statistischer Methoden (Technische Großzahlforschung). Vorträge, gehalten am Außeninstitut der Technischen Hochschule Berlin im Wintersemester 1928/29, Berlin 1930.

Da aber die mittlere Abweichung nach Formel (4) von Art. 107 gleich 52,6 ist, so liegt die vorstehende Differenz sogar unter der einfachen mittleren Abweichung und kann daher ganz wohl einer zufälligen Störung des normalen Verlaufs zugeschrieben werden.

Anders steht es um die Klasse 60 bis 61, die schon äußerlich durch den schroffen Abfall der Häufigkeit gegenüber der Nachbarklasse 61 bis 62 auffällt. Die Zahlwerte ξ der Klassengrenzen sind 2,234, 2,625, die zugehörigen Tafelwerte 0,98724, 0,99566, daraus die gerechnete relative Häufigkeit 0,00842 und die absolute 0,00842 25878 = 218, was gegen die beobachtete um 168 größer ist.

Die mittlere Abweichung der Klasse beträgt aber nach Formel (4) von Art. 107 14,7, 168 ist mehr als das 11fache davon; die Abweichung kann nicht als eine zufällige gelten, dürfte vielmehr einer besonderen Ursache entsprungen sein.

126. Wenn eine Verteilung der Normalkurve entspricht, so besteht zwischen den einzelnen Streuungsmaßen Proportionalität. Wir suchen die Proportionalitätsfaktoren der durchschnittlichen Abweichung ϑ und des Quartils Q in Bezug auf die mittlere Abweichung. Da bei unserer Rechnung μ als Einheit gilt, so ergibt sich der erstere Proportionalitätsfaktor als Wert des Integrals

$$\frac{2}{\sqrt{2\pi}}\int_0^\infty \xi e^{-\frac{\xi^2}{2}} d\xi = \frac{2}{\sqrt{2\pi}}\left| e^{-\frac{\xi^2}{2}}\right|_\infty^0 = \sqrt{\frac{2}{\pi}} = 0{,}79788\ldots,$$

so daß

$$\vartheta = \sqrt{\frac{2}{\pi}}\,\mu = 0{,}79788\ldots\mu; \tag{13}$$

der letztere Proportionalitätsfaktor als Auflösung der Gleichung

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\chi} e^{-\frac{\xi^2}{2}} d\xi = \frac{3}{4}$$

nach χ. Die Tab. 67 (Art. 124) liefert bei linearer Interpolation

$$\chi = 0{,}65 + \frac{786}{1590}\cdot 0{,}05 = 0{,}675,$$

bei schärferer Rechnung aus einer größeren Tafel findet man

$$\chi = 0{,}67449,$$

so daß

$$Q = \chi\mu = 0{,}67449\,\mu. \tag{14}$$

Auf diesen Formeln beruhen die Näherungsregeln, die durchschnittliche Abweichung betrage $\frac{4}{5}$, das Quartil oder die wahrscheinliche Abweichung $\frac{2}{3}$ der mittleren Abweichung.

In der Fehlertheorie heißt ϑ der durchschnittliche, Q der wahrscheinliche Fehler einer Beobachtung. Der Grund, warum manche den letzteren als Genauigkeitsmaß (als Streuungsmaß) vorziehen, liegt darin, daß man ihm die leichtverständliche Erklärung geben kann, es falle die Hälfte der Fälle in das Intervall $(-Q, Q)$, die andere Hälfte darüber hinaus.

Für die Intervalle $(-\mu, +\mu)$, $(-\vartheta, +\vartheta)$ gibt es auch eine feste, aber nicht so einfache Einteilung. Man findet sie mittels der zu $\xi = 1$, $\xi = \sqrt{\frac{2}{\pi}}$ gehörigen Werte von Ψ; den ersten gibt die Tabelle unmittelbar $= 0{,}84134$, den zweiten durch Interpolation $= 0{,}78751$; daraus leiten sich die verlangten Häufigkeiten ab:

$$2 \cdot 0{,}84134 - 1 = 0{,}6827$$
$$2 \cdot 0{,}78751 - 1 = 0{,}5750;$$

man kann also sagen, in das Intervall $(-\mu, \mu)$ fallen rund 68, darüber hinaus 32, in das Intervall $(-\vartheta, \vartheta)$ rund 58, darüber hinaus 42% aller Fälle.

Es ist zu beachten, daß diese Überführungen der Streuungsmaße ineinander nur bei einer normalen Verteilung Geltung haben. Es ist vorzuziehen bei der mittleren Abweichung zu bleiben, weil ihre Bedeutung unabhängig ist von der Art der Verteilung.

Setzt man, um auf einen speziellen Fall die Probe zu machen, bei den amerikanischen Rekruten $\mu = 2{,}55$ und leitet daraus gemäß den Formeln (13) und (14) ϑ und Q ab, so findet man $\vartheta = 2{,}03$, $Q = 1{,}72$; bei linearer Interpolation ergeben sich für die Intervalle (64,15, 69,25), (64,67, 68,73), (64,98, 68,42) unter Zugrundelegung der Reihe der beobachteten Häufigkeiten die folgenden Häufigkeiten

17378 14443 12536

und daraus durch Division mit 25878 die relativen

0,672 0,558 0,484,

während sie der Theorie nach

0,683 0,575 0,500

zu betragen hätten. Auch darin liegt ein Beweis dafür, daß die Verteilung der Körperhöhen den Hauptzug des Normalen trägt.

127. Bei der Ableitung der Funktion z, welche zur normalen Häufigkeitskurve führt, aus der Binomialformel war es eine wesentliche Voraussetzung, daß nicht bloß n, die Anzahl der Versuche oder Beobachtungen, sondern auch die Produkte np, nq große Zahlen seien, daß also keine der beiden Wahrscheinlichkeiten p, q eine sehr kleine, der Null naheliegende Zahl sei. Daraus folgt, daß sich die Verteilung sehr selten — im Vergleich zur Gesamtzahl der Fälle — auftretender Ereignisse nicht wird unter das Gesetz der Normalkurve bringen lassen. Ereignisse solcher Art sind unschwer nachzuweisen: Blindgeburten unter der Gesamt-

heit der Geborenen eines Gebiets, Kinderselbstmorde, Frauenselbstmorde innerhalb einer Bevölkerung, tödlich verlaufende Betriebsunfälle bei verschiedenen Betriebsarten, Tötungen durch ganz besondere Ursachen, wie durch Hufschlag, durch Explosion u. ä.

Poisson hatte bereits auf diese Ausnahme hingewiesen [1]), hat ihre analytische Behandlung eingeleitet, ohne sie jedoch bis zu einer Anwendung weiter zu verfolgen. Neu aufgenommen wurde die Frage nach dem Verhalten solcher in langen Beobachtungsreihen selten vorkommenden Ereignisse von L. v. Bortkiewicz [2]), der ihr eine eingehende Untersuchung zuteil werden ließ und Beobachtungsmaterial beibrachte, an welchem die Stichhaltigkeit der Theorie geprüft werden konnte. Als Seitenstück zu Poissons „Gesetz der großen Zahlen" schlug er für den Ausnahmefall die Bezeichnung „Gesetz der kleinen Zahlen" vor; während dort unter den großen Zahlen die Umfänge der Versuchsreihen gemeint sind, ist hier mit den kleinen Zahlen auf die Wiederholungszahlen der seltenen Ereignisse angespielt. Das Wesentliche an der Sache ist, daß auch solche Ereignisse, entgegen der ursprünglichen Meinung, sie würden sich wegen ihrer besonderen Verursachung einem Zufallsgesetz nicht fügen, im allgemeinen, mitunter überraschend gut, jenem mathematischen Gesetz sich anpassen, das aus der Binomialformel unter der jetzt geltenden Voraussetzung einer sehr kleinen Wahrscheinlichkeit hervorgeht.

Angenommen also, von den beiden Wahrscheinlichkeiten p, q sei die zweite, q, sehr klein, die Anzahl n der Versuche aber in einer noch näher festzusetzenden Weise sehr groß. Die zu p und q gehörigen Ereignisse sollen beziehungsweise E und F heißen.

Man kann erstens so vorgehen, daß man unter der eben gemachten Voraussetzung einen Näherungsausdruck sucht für die in Art. 111 entwickelte Formel der Bernoullischen Verteilung

$$P_m = \frac{n!}{(n-m)!\,m!}\, p^{n-m}\, q^m,$$

d. i. für die Wahrscheinlichkeit, daß sich in n Versuchen das Ereignis F m mal einstellt; dabei kommen nur verhältnismäßig kleine Werte von m in Betracht, nämlich Werte von der Ordnung der Zahl nq, von der wir annehmen, daß sie trotz der Kleinheit von q infolge eines entsprechend großen n einen endlichen Wert hat, der mit λ bezeichnet werden möge, so daß man für q auch schreiben kann $\frac{\lambda}{n}$. Läßt man in der Formel $m!$ mit seinem strengen Wert stehen, ersetzt aber $n!$ und $(n-m)!$ durch ihre Ausdrücke nach der Stirlingschen Formel (Art. 113 (2)), so wird nach entsprechender Kürzung

$$P_m = \frac{(n-m)^n}{\left(1-\frac{m}{n}\right)^{n+\frac{1}{2}}}\, e^{-m}\, p^{n-m}\, q^m;$$

[1]) Poisson, Recherches sur la probabilité des jugements, Paris 1837, in der deutschen Übersetzung von E. H. Schnuse, 1841, § 81, S. 171.

[2]) L. v. Bortkiewicz, Das Gesetz der kleinen Zahlen, Leipzig 1898.

bei sehr großem n unterscheidet sich $\left(1-\frac{m}{n}\right)^{n+\frac{1}{2}}$ sehr wenig von $\left(1-\frac{m}{n}\right)^n$ und dies wieder sehr wenig von e^{-m}; des weitern $p^{n-m}=(1-q)^{n-m}$ sehr wenig von $e^{-(n-m)q}$, da einerseits

$$(1-q)^{n-m}=1-(n-m)\,q+\frac{(n-m)\;(n-m-1)}{1\cdot 2}\,q^2-\cdots,$$

andererseits

$$e^{-(n-m)\,q}=1-(n-m)\,q+\frac{(n-m)^2}{1\cdot 2}\,q^2-\cdots;$$

führt man diese Näherungen ein, so wird weiter

$$P_m=\frac{(n-m)^m}{m!}\,q^m\,e^{-(n-m)\,q};$$

ersetzt man aber $(n-m)\,q$ durch das nur wenig davon abweichende $n\,q$, wofür der Buchstabe λ eingeführt worden ist, so kommt man schließlich zu dem Näherungsausdruck

$$P_m=\frac{\lambda^m\,e^{-\lambda}}{m!}. \tag{15}$$

für die Wahrscheinlichkeit, daß das Ereignis F in n Versuchen m mal eintrifft.

Ein anderer Weg, der die Zulässigkeit des vorstehenden Näherungsverfahrens deutlicher macht, ist der folgende. Wir fragen nach der Wahrscheinlichkeit, daß das Ereignis F in n Versuchen höchstens m mal, das Ereignis E also mindestens $(n-m)$ mal sich zutragen werde.

Das kann erstens in der Weise geschehen, daß sich E in den ersten $n-m$ Versuchen durchwegs einstellt, wofür die Wahrscheinlichkeit

$$p^{n-m}$$

besteht; in den weiteren m Versuchen kann sich dann F höchstens m mal zutragen.

Oder zweitens derart, daß E in den ersten $(n-m+1)$ Versuchen $(n-m)$ mal eintrifft, jedoch so, daß es auch an letzter Stelle erscheint, denn sonst wäre der vorige Fall eingetreten; die Wahrscheinlichkeit hierfür ist

$$(n-m)\,p^{n-m}\,q,$$

weil $(n-m)$ Plätze für das Ereignis F zur Verfügung stehen.

Oder drittens in der Weise, daß in den ersten $(n-m+2)$ Versuchen das Ereignis E sich $(n-m)$ mal einstellt, jedoch so, daß es auch an letzter Stelle erscheint, weil sonst einer der zwei früheren Fälle eingetreten sein müßte; diesem Sachverhalt entspricht die Wahrscheinlichkeit

$$\frac{(n-m+1)\;(n-m)}{1\cdot 2}\,p^{n-m}\,q^2$$

entsprechend den $\frac{(n-m+1)\,(n-m)}{1\cdot 2}$ Platzpaaren, die dem F zur Verfügung stehen.

So fortfahrend kommt man schließlich zu dem Falle, wo E in den n Versuchen $(n-m)$ mal, darunter auch an letzter Stelle, und $F\,m$ mal erscheint, wofür die Wahrscheinlichkeit

$$\frac{(n-1)(n-2)\ldots(n-m)}{1\cdot 2\ldots m}\,p^{n-m}q^m$$

besteht. Hiernach stellt sich die erfragte Wahrscheinlichkeit Π_m auf:

$$\left.\begin{aligned}\Pi_m = p^{n-m}\Big[1 + (n-m)\,q + \frac{(n-m)(n-m+1)}{1\cdot 2}\,q^2 + \\ + \frac{(n-m)(n-m+1)(n-m+2)}{1\cdot 2\cdot 3}\,q^3 + \cdots + \frac{(n-m)(n-m+1)\ldots(n-1)}{1\cdot 2\ldots m}\,q^m\Big].\end{aligned}\right\} \quad (16)$$

Es gibt noch eine andere Form ihrer Darstellung als jener Teil der Entwicklung von $(p+q)^n$, in welchem die Exponenten von q von 0 bis m, jene von p also von n bis $n-m$ gehen, also

$$\left.\begin{aligned}\Pi_m = p^n + n\,p^{n-1}q + \frac{n(n-1)}{1\cdot 2}\,p^{n-2}q^2 + \cdots + \\ + \frac{n(n-1)\ldots(n-m+1)}{1\cdot 2\ldots m}\,p^{n-m}q^m.\end{aligned}\right\} \quad (17)$$

Die Übereinstimmung der beiden Ausdrücke, die hier aus wahrscheinlichkeitstheoretischen Gründen unmittelbar hervorgeht, ist algebraisch so zu erweisen: man ersetze in (17)

$$\begin{aligned} p^n &\text{ durch } p^{n-m}(1-q)^m \\ p^{n-1} &\quad ,, \quad p^{n-m}(1-q)^{m-1} \\ &\ldots \\ p^{n-m+1} &\quad ,, \quad p^{n-m}(1-q), \end{aligned}$$

entwickle die Binompotenzen und ordne steigend nach q.

Führt man an der Formel (16) die Näherungsrechnung aus, indem man $(n-m)q$ durch λ, $p^{n-m} = (1-q)^{n-m}$ durch $e^{-\lambda}$ ersetzt und λq neben λ^2, $2\lambda q^2$, $3\lambda^2 q$ neben λ^3 u. s. w. wegläßt, so ergibt sich für sehr kleine q und sehr große n die Näherungsformel

$$\Pi_m = e^{-\lambda}\left[1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots + \frac{\lambda^m}{m!}\right]. \quad (18)$$

Zufolge der Bedeutung von Π_m ist $\Pi_m - \Pi_{m-1}$, also $\frac{\lambda^m}{m!}e^{-\lambda}$, die Wahrscheinlichkeit, daß das Ereignis F in n Versuchen gerade m mal eintrifft, in Übereinstimmung mit der Formel (15).

Eine andere Probe auf die Haltbarkeit der Formeln ergibt sich, wenn man in (18) m durch n ersetzt; dann sollte

$$\Pi_m = e^{-\lambda}\left[1 + \lambda + \frac{\lambda^2}{2!} + \quad + \frac{\lambda^n}{n!}\right] \tag{19}$$

den Wert 1 annehmen, weil es gewiß ist, daß in n Versuchen F höchstens n-mal eintrifft; in der Tat ist der Limes des Klammerinhalts für $n \to \infty$ gleich e^λ und für sehr große n sehr wenig davon verschieden.

Die Glieder des entwickelten Produktes (19) geben die relative Verteilung der Werte 0, 1, 2, ... n von m an. Multipliziert man sie der Reihe nach mit eben diesen Zahlen, so ergibt sich in der Summe dieser Produkte der Mittelwert von m, nämlich

$$\left.\begin{aligned} & e^{-\lambda}\left[\lambda + \lambda^2 + \frac{\lambda^3}{2!} + \cdots + \frac{\lambda^n}{(n-1)!}\right] = \\ & = \lambda e^{-\lambda}\left[1 + \lambda + \frac{\lambda^2}{2!} + \cdots + \frac{\lambda^{n-1}}{(n-1)!}\right] = \lambda, \end{aligned}\right\} \tag{20}$$

weil bei sehr großem n der Klammerausdruck nur sehr wenig von e^λ abweicht; multipliziert man weiter die Glieder von (20) der Reihe nach mit 1, 2, ... n, bildet wiederum die Summe und subtrahiert davon das Quadrat von (20), d. i. λ^2, so ergibt sich das Quadrat der mittleren Abweichung der Wiederholungszahl m von ihrem Mittelwert, also

$$\left.\begin{aligned} \mu^2 = & \; e^{-\lambda}\left[\lambda + 2\lambda^2 + \frac{3\lambda^3}{2!} + \cdots + \frac{n\lambda^n}{(n-1)!}\right] - \lambda^2 = \\ & = \lambda e^{-\lambda}\left[1 + 2\lambda + \frac{3\lambda^2}{2!} + \quad + \frac{n\lambda^{n-1}}{(n-1)!}\right] - \lambda^2 = \\ & = \lambda e^{-\lambda}\left[1 + \lambda + \frac{\lambda^2}{2!} + \cdots + \frac{\lambda^{n-1}}{(n-1)!}\right] + \\ & + \lambda^2 e^{-\lambda}\left[1 + \lambda + \frac{\lambda^2}{2!} + \cdots + \frac{\lambda^{n-2}}{(n-2)!}\right] - \lambda^2 = \\ & = \lambda + \lambda^2 - \lambda^2 = \lambda \end{aligned}\right\} \tag{21}$$

wiederum, weil die Klammerinhalte bei sehr großem n nur sehr wenig von e^λ sich unterscheiden.

Das erste dieser beiden Resultate steht im Einklang mit dem a priori zu erwartenden Werte von m, der ja $nq = \lambda$ ist. Das zweite Ergebnis ist für unseren Fall bemerkenswert und stimmt auch zu früheren Ergebnissen; denn wir fanden das Quadrat der mittleren Abweichung der Wiederholungszahlen von E und F gleich npq, also auch gleich $nq - nq^2$, und das ist bei sehr kleinem q und endlichem nq nur sehr wenig verschieden von $nq = \lambda$.

Die mathematische Grundlegung des Gesetzes der kleinen Zahlen, sein Name und die Prioritätsfrage waren Gegenstand einer Kritik seitens Lucy Whitaker,

einer Schülerin und Mitarbeiterin Pearsons, gegen die sich Bortkiewicz mit einer Gegenschrift[1]) verteidigt hat. Die Kritik greift an dem Punkte an, daß in der Formel (15) q und n nicht einzeln, sondern nur in Form des Produktes vorkommen, für welches Produkt bei der praktischen Verfolgung der empirische Mittelwert von m genommen wird. L. Whitaker meint daher, man solle n, q aus der Gleichstellung des empirisch bestimmten Mittelwertes von m mit dem theoretischen nq und aus der Gleichstellung der empirisch aus der Verteilung abgeleiteten mittleren Abweichung mit ihrem theoretischen Wert $\sqrt{npq}$ nachträglich berechnen und darnach erst entscheiden, ob auf den betreffenden Fall die Formeln anwendbar seien. Die Durchführung dieses Gedankens muß notwendig zu Ungereimtheiten führen, da man die dem Zufall unterworfenen und von ihm entstellten empirischen Werte zur Grundlage einer Berechnung macht; kleine Einflüsse können hierbei große Wirkungen hervorbringen. Was die Namengebung betrifft, so besteht zwischen „Gesetz der kleinen Zahlen“ und „Gesetz der großen Zahlen“ eine nicht zu leugnende Inkongruenz; in beiden Fällen handelt es sich um große Gesamtzahlen von Einzelerfahrungen, also um die Aussage, daß die Ereignisse in langen Reihen von Beobachtungen nahe im Verhältnis ihrer Wahrscheinlichkeiten sich zutragen, ob diese Wahrscheinlichkeiten groß oder klein sind. In der Bezeichnung „Gesetz der kleinen Zahlen“ wird aber etwas anderes hervorgehoben, nämlich, daß es sich um Ereignisse mit (selbst in sehr langen Reihen) kleiner Wiederholungszahl, also um selten eintreffende Ereignisse handelt. In der Prioritätsfrage muß festgestellt werden, daß Poisson auf eine Bewährung des Gesetzes der großen Zahlen bei kleinen Ereigniszahlen nicht einging, sondern bei der bloßen Ableitung der Grenzformel (15) stehen blieb. Die sachliche Begründung gab Bortkiewicz.

Zum besseren Verständnis der Sache sei die Tab. 68 eingefügt, die die Verteilungen (19) für eine Reihe von Werten des λ zur Anschauung bringt. Sie ist ein Auszug aus der ausführlichen Tafel, die Bortkiewicz[2]) gegeben hat und die nach Zehnteln fortschreitend von $\lambda = 0{,}1$ bis $\lambda = 10$ geht. In der Tab. 68 sind im besonderen die Werte des λ berücksichtigt, die für die Durchrechnung der folgenden Beispiele benötigt werden.

Wie man bemerkt, ist die Verteilung bei Werten von λ unter und bis 1 einseitig, wird dann stark asymmetrisch, doch rückt der Gipfelpunkt mit wachsendem λ immer weiter nach rechts. Bei $m = \lambda - 1$ und $m = \lambda$ ergeben sich zwei gleiche Glieder.

Werden beispielsweise über ein Ereignis, dessen Wahrscheinlichkeit 0,0004 ist, 100 Serien von je 10000 Versuchen angestellt, so sind laut der letzten Kolonne

2 7 15 20 20 16 10 6 3 1

Serien mit beziehungsweise

0 1 2 3 4 5 6 7 8 9-

maligem Eintreffen des betreffenden Ereignisses zu erwarten.

[1]) L. v. Bortkiewicz, Realismus und Formalismus in der mathematischen Statistik. Allgemeines Statistisches Archiv, Bd. IX, 1915, S. 225—256.

[2]) L. v. Bortkiewicz, Das Gesetz der kleinen Zahlen. Leipzig 1898, S. 49 u. f.

Tab. 68. Tafel der Werte von $\frac{\lambda^m e^{-\lambda}}{m!}$.

m	$\lambda = 0,1$	$\lambda = 0,5$	$\lambda = 0,6$	$\lambda = 0,7$	$\lambda = 1,0$	$\lambda = 2,0$	$\lambda = 3,0$	$\lambda = 4,0$	$\lambda = 5,0$	$\lambda = 5,1$
0	0,9048	0,6065	0,5488	0,4966	0,3679	0,1353	0,0498	0,0183	0,0067	0,0061
1	0,0905	0,3033	0,3293	0,3476	0,3679	0,2707	0,1494	0,0733	0,0337	0,0311
2	0,0045	0,0758	0,0988	0,1217	0,1839	0,2707	0,2240	0,1465	0,0842	0,0793
3	0,0002	0,0126	0,0198	0,0284	0,0613	0,1804	0,2240	0,1954	0,1404	0,1348
4	.	0,0016	0,0030	0,0050	0,0153	0,0902	0,1680	0,1954	0,1755	0,1719
5	.	0,0002	0,0004	0,0007	0,0031	0,0361	0,1008	0,1563	0,1755	0,1753
6	.	.	.	0,0001	0,0005	0,0120	0,0504	0,1042	0,1462	0,1490
7	.	.	.	.	0,0001	0,0034	0,0216	0,0595	0,1044	0,1086
8	.	.	.	.	.	0,0009	0,0081	0,0298	0,0653	0,0692
9	.	.	.	.	.	0,0002	0,0027	0,0132	0,0363	0,0392
10	.	.	.	.	.	.	0,0007	0,0053	0,0181	0,0200
11	.	.	.	.	.	.	0,0001	0,0019	0,0082	0,0093
12	.	.	.	.	.	.	.	0,0006	0,0034	0,0039
13	.	.	.	.	.	.	.	0,0002	0,0013	0,0015
14	.	.	.	.	.	.		0,0001	0,0005	0,0006
15	.	.	.	.	.	.	.	.	0,0002	0,0002
16	.	.	.	.	.	.	.	.	.	0,0001

128. Beispiele.

1) Die Statistik der weiblichen Selbstmorde[1]) im Alter von 30 bis 60 Jahren in den 27 sächsischen Amtshauptmannschaften während der Jahre 1932 bis 1934 ergab $27 \cdot 3 = 81$ Ereigniszahlen, die von 0 (kein weiblicher Selbstmord) bis 15 (15 solche Selbstmorde in einem Jahr und einem Gebiet) gingen. Bei der Wertung dieses Materials darf nicht übersehen werden, daß die einzelnen Ereigniszahlen nicht gleichwertig sind, weil aus Gebieten verschiedener Größe stammend. Das zeigt sich auch bei ihrer zeitlichen und örtlichen Gliederung: in den kleinen Gebieten überwiegen die kleinen Ereigniszahlen und fehlen die großen, in den größeren herrscht das umgekehrte Verhalten.

Die folgende Tab. 69 gibt die Verteilung der 81 Ereigniszahlen (Spalte z) auf die zur Beobachtung gelangten Werte von 0 bis 15 (Spalte m); die Summe der mit mz überschriebenen Kolonne bedeutet die Zahl aller Selbstmorde, die somit 406 betrug. Die weiter angefügten Kolonnen dienen der Bestimmung des arithmetischen Mittels und der mittleren Abweichung nach der Summenmethode; die bezüglichen Rechnungen finden sich unterhalb der Tabelle.

Tab. 69. Weibliche Selbstmorde.

m	z	mz		
0	2	0	.	.
1	2	2	79	327
2	10	20	77	829
3	12	36	67	250
4	14	56	55	183
5	12	60	41	128
6	9	54	29	87
7	10	70	20	58
8	3	24	10	38
9	2	18	7	28
10	—	—	5	21
11	—	—	5	16
12	1	12	5	11
13	3	39	4	6
14	—	—	1	2
15	1	15	1	1
	81	406		

$$\lambda = \frac{79 + 327}{81} = 5{,}01 \qquad \mu^2 = \frac{79 + 3 \cdot 327 + 2 \cdot 829}{81} - 5{,}01^2 = 8{,}4555.$$

[1]) Zeitschrift des Sächsischen Statistischen Landesamtes 1936, S. 78.

m	z berechnet	z beobachtet
0	0,5	2
1	2,7	2
2	6,8	10
3	11,3	12
4	14,2	14
5	14,2	12
6	11,9	9
7	8,5	10
8	5,3	3
9	3,0	2
10	1,5	—
11	0,7	—
12	0,3	1
13	0,1	3
14	0,0	—
15	0,0	1
	81,0	81

Für λ ist der Theorie nach das arithmetische Mittel der m zu nehmen; ihr zufolge sollte μ^2 mit λ übereinstimmen; die große Abweichung dürfte aus der oben erwähnten Ungleichwertigkeit der Zahlen zu erklären sein. Man hat also zwei Bestimmungen der mittleren Abweichung:

die theoretische: ..... $\sqrt{5{,}01} = 2{,}24$

die empirische: ...... $\sqrt{8{,}4555} = 2{,}91$,

die erheblich auseinandergehen.

Mit dem gefundenen λ ist dann auf Grund der Tab. 68 die theoretische Verteilung gerechnet und neben die beobachtete gestellt worden; dies gibt das nebenstehende Bild.

Beide Reihen zeigen wohl den gleichen Grundzug, gehen aber im einzelnen verschiedentlich auseinander.

2) In 10 preußischen Armeekorps sind in den 20 Jahren 1875 bis 1894 im ganzen 122 Mann durch Hufschlag getötet worden[1]). Die Verteilung der $10 \times 20 = 200$ Ereigniszahlen, die von 0 bis 4 gingen, war die folgende:

Tab. 70. Militärsterbefälle durch Hufschlag.

m	z	$m z$		
0	109	0	.	.
1	65	65	91	31
2	22	44	26	6
3	3	9	4	5
4	1	4	1	1
	200	122		

$$\lambda = \frac{91 + 31}{200} = 0{,}61 \qquad \mu^2 = \frac{91 + 3 \cdot 31 + 2 \cdot 6}{200} - 0{,}61^2 = 0{,}61.$$

[1]) L. v. Bortkiewicz, Das Gesetz der kleinen Zahlen, Leipzig 1898, S. 23 u. f.

Die Theorie findet hier volle Bestätigung, indem der Mittelwert mit dem Quadrat der mittleren Abweichung genau übereinstimmt, was jedoch nur als eine seltene Ausnahme aufzufassen ist. Die theoretische Verteilung ergibt sich, indem man die 200fachen Glieder von (19) für $\lambda = 0{,}61$ bis $m = 4$ rechnet; sie ist nachstehend der wirklich beobachteten gegenübergestellt und zeigt eine außergewöhnliche Übereinstimmung, wie sie bei solchen Materien selten zu treffen sein wird.

m	z berechnet	z beobachtet
0	108,7	109
1	66,3	65
2	20,2	22
3	4,1	3
4	0,6	1
5 und mehr	0,1	—
	200,0	200

Das Gesetz der kleinen Zahlen spielt auch in der Verkehrsstatistik eine Rolle, und zwar bei der Auswertung der Ergebnisse von Verkehrskontrollen Vgl. hierzu H. Kellerer, Mathematische Probleme in der Verkehrsstatistik. Archiv für mathematische Wirtschafts- und Sozialforschung. Bd. I, Heft 1, S. 50 u. f.

129. In Art. 107 sind in den Formeln (3), (4) und (5) die Ausdrücke für den Mittelwert p und für die Streuung μ der unbekannten Wahrscheinlichkeit η angegeben worden. Es handelt sich hierbei um den Fall, daß p auf empirischer Basis bestimmt wird. In dem anderen Falle, in dem die Grundwahrscheinlichkeit logisch erschlossen werden kann, gelten die Formeln (1) und (2) von Art. 106.

Zur Herleitung der Formeln (3), (4) und (5) von Art. 107 gehen wir mit Bezug auf Formel (1b) in Art. 111 von der folgenden Gleichung aus

$$w(\eta) = C f(\eta)\, \eta^m (1-\eta)^n \tag{22}$$

Hierin bedeuten $w(\eta)$ die Wahrscheinlichkeit dafür, daß die Wahrscheinlichkeit des Erfolgs gerade η ist; $f(\eta)$ die entsprechende Anfangswahrscheinlichkeit; m die Anzahl der empirisch bestimmten Erfolge bei n Versuchen. Für die Anfangswahrscheinlichkeit $f(\eta)$ setzen wir die Konstante C_0. Das Produkt der Konstanten $C \cdot C_0 = C'$ ergibt sich aus der Festsetzung

$$\int w(\eta)\, d\eta = 1 \tag{23}$$

zu

$$C' = \frac{}{\int^{1} \eta^m (1-\eta)^{n-m}\, d\eta}$$

Wir erhalten nach der Integralformel

$$\int_0^1 x^k (1-x)^l \, dx = \frac{k! \; l!}{(k+l+1)!} \tag{24}$$

$$C' = \frac{(n+1)!}{m! \; (n-m)!} = (n+1)\binom{n}{m},$$

und somit ist

$$w(\eta) = \frac{(n+1)!}{m! \; (n-m)!} \eta^m (1-\eta)^{n-m}. \tag{25}$$

Diese Gleichung kennzeichnet den Inhalt des Bayesschen Problems[1]).

Den Mittelwert p der unbekannten Wahrscheinlichkeit η berechnen wir in folgender Weise:

$$p = \int_0^1 \eta \, w(\eta) \, d\eta \tag{26}$$

$$p = \frac{(n+1)!}{m! \; (n-m)!} \int_0^1 \eta^{m+1} (1-\eta)^{n-m} \, d\eta.$$

Mittels der Integralformel (24) stellt sich p auf

$$p = \frac{(n+1)!}{m! \; (n-m)!} \cdot \frac{(m+1)! \; (n-m)!}{(n+2)!}$$

$$p = \frac{m+1}{n+2} \tag{27}$$

Damit haben wir die Gleichung (3) in Art. 107 hergeleitet.

Die Gleichung (5), die sich auf die Streuung $\mu_{(1)}$ bezieht, bestimmen wir wie folgt:

$$\mu_{(1)}^2 = \int^1 (\eta - p)^2 \, w(\eta) \, d\eta. \tag{28}$$

Es ist unter Berücksichtigung von (23) und (26)

$$\mu_{(1)}^2 = \int_0^1 \eta^2 \, w(\eta) \, d\eta - p^2.$$

Hieraus folgt sofort

$$\mu_{(1)}^2 = \frac{(n+1)!}{m! \; (n-m)!} \int^1 \eta^{m+2} (1-\eta)^{n-m} \, d\eta - p^2$$

[1]) Vgl. Thomas Bayes, Versuch zur Lösung eines Problems der Wahrscheinlichkeitsrechnung. Ostwalds Klassiker der exakten Wissenschaften Nr. 169.

und unter Anwendung der Integralformel (24)

$$\mu_{(1)}^{2} = \frac{(n+1)!}{m!\,(n-m)!} \cdot \frac{(m+2)!\,(n-m)!}{(n+3)!} - p^2 = \frac{(m+1)\,(m+2)}{(n+2)\,(n+3)} - p^2$$

$$\mu_{(1)}^{2} = \frac{p\,(1-p)}{n+3}. \tag{29}$$

Damit ist auch die Gleichung (5) in Art. 107 entwickelt. Aus Gleichung (29) folgt durch eine einfache Überlegung die Gleichung (4) in Art. 107.

130. Zum Schlusse dieses Paragraphen sei noch die Frage nach der Verläßlichkeit in der Bestimmung der Perzentilen erörtert.

Denken wir uns ein stetiges Kollektiv von unendlichem Umfang und bezeichnen mit $X(v)$ jenen Argumentwert, der das Kollektiv im Verhältnis $v:(1-v)$ teilt, so daß der Anteil v über, der Anteil $1-v$ unter $X(v)$ liegt.

Denken wir uns weiter dem Kollektiv eine ständig sich erweiternde Probe entnommen, so wird das Teilungsverhältnis, welches der Argumentwert $X(v)$ in ihr bewirkt, sich ständig dem Verhältnis $v:(1-v)$ nähern und wenn die Probe den Umfang n erlangt hat, wird die mittlere Abweichung davon $\sqrt{\frac{v(1-v)}{n}}$ sein.

Andererseits, wenn bei Erreichung dieses Umfangs die wirkliche Verhältniszahl, statt v zu sein, $v+h$ ist, so entspricht ihr ein anderer Argumentwert, nämlich $X(v+h)$, der sich von dem früheren um eine Größe δ unterscheidet, so daß

$$\delta = X(v+h) - X(v) = \frac{dX}{dv}h; \tag{30}$$

bei dieser Darstellung ist die Voraussetzung genügender Kleinheit von h gemacht, um Glieder höheren Grades, die sich bei der Entwicklung in die Taylorsche Reihe ergeben, weglassen zu dürfen.

Es handelt sich nun um die Bestimmung des Differentialquotienten $\frac{dX}{dv}$, dessen Wert natürlich von dem Verteilungsgesetz abhängen wird. Wir denken uns dieses durch die Häufigkeitskurve, Fig. 37, dargestellt und als Streuungsmaß die mittlere Abweichung μ bestimmt. Dann liegt die Sache so, daß die zur Abszisse $X(v)$ gehörige Ordinate z_v der Kurve deren Fläche F in die Teile vF und $(1-v)F$ teilt; nennen wir den ersten, über $X(v)$ liegenden Teil $F(v)$, so ist also

Fig. 37. Zur Fehlerbestimmung der Perzentilen.

$$F(v) = vF,$$

woraus durch Differentiation nach v erhalten wird

$$\frac{dF(v)}{dX} \cdot \frac{dX}{dv} = F,$$

und da das Flächendifferential $dF(v) = z_v\,dX$, also die Endordinate $z_v = \frac{dF(v)}{dX}$ ist, so folgt weiter

$$z_v \frac{dX}{dv} = F;$$

nimmt man die auf μ als Basis bezogene Fläche der ganzen Kurve als Einheit, so folgt ferner

$$\frac{z_v}{\mu} \frac{dX}{dv} = 1$$

und daraus ergibt sich

$$\frac{dX}{dv} = \frac{\mu}{z_n}.$$

Mithin besteht zwischen den beiden Inkrementen h und δ der Zusammenhang

$$\delta = \frac{\mu}{z_v} h$$

und daraus folgt für die mittleren Abweichungen, einerseits μ_v im Teilungsverhältnis, andererseits μ_X in der Lage des Teilungspunktes $X(v)$, die Beziehung

$$\mu_X = \frac{\mu}{z_n} \mu_v;$$

für μ_v ist aber oben der Ausdruck $\sqrt{\frac{v(1-v)}{n}}$ gefunden worden; daher ist endgültig

$$\mu_X = \frac{\mu}{z_v} \sqrt{\frac{v(1-v)}{n}} \tag{31}$$

Bemerkenswert ist das Auftreten von z_v in dieser Formel. Es zeigt, daß die Unsicherheit einer Perzentile um so größer ist, je kleiner die Häufigkeit an der Stelle ist, in welche sie fällt, und umgekehrt.

Bei analytisch gegebener Häufigkeitskurve läßt sich die Formel vollständig durchführen. Es soll dies gezeigt werden für den Fall einer normalen Verteilung.

Um z_v zu finden, muß man $X(v)$ kennen; dieses aber bestimmt sich aus dem Ansatz

$$\frac{1}{\mu\sqrt{2\pi}} \int_0^{X(v)} e^{-\frac{x^2}{2\mu^2}} dx = \frac{1}{2} - v$$

oder aus

$$\frac{2}{\sqrt{\pi}} \int_0^{\frac{X(v)}{\mu\sqrt{2}}} e^{-t^2} dt = 1 - 2v. \tag{32}$$

Für die praktische Anwendung kommen hauptsächlich folgende Teilungen in Betracht: der Zentralwert, entsprechend $v = \frac{1}{2}$; die Dezilen, entsprechend $v = \frac{1}{10}, \frac{2}{10}, \frac{3}{10}, \frac{4}{10}$; die Quartilen, entsprechend $v = \frac{1}{4}$. Durch Einsetzung dieser Werte in (32) und Interpolation in der Tafel des Integrals findet man

$$X\left(\frac{1}{2}\right) = 0$$

$$X\left(\frac{1}{10}\right) = 0{,}90622\,\mu\sqrt{2} = 1{,}28159\,\mu$$

$$X\left(\frac{2}{10}\right) = 0{,}59513\,\mu\sqrt{2} = 0{,}84163\,\mu$$

$$X\left(\frac{3}{10}\right) = 0{,}37081\,\mu\sqrt{2} = 0{,}52440\,\mu$$

$$X\left(\frac{4}{10}\right) = 0{,}17914\,\mu\sqrt{2} = 0{,}25334\,\mu$$

$$X\left(\frac{1}{4}\right) = 0{,}47695\,\mu\sqrt{2} = 0{,}67451\,\mu.$$

Daraus berechnen sich mittels der Gleichung

$$z = \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}},$$

wenn man für ξ die obigen Zahlenkoeffizienten 0, 1,28159 u. s. w. einträgt,

$$z_{\frac{1}{2}} = 0{,}3989$$

$$z_{\frac{1}{10}} = 0{,}1756$$

$$z_{\frac{2}{10}} = 0{,}2799$$

$$z_{\frac{3}{10}} = 0{,}3475$$

$$z_{\frac{4}{10}} = 0{,}3859$$

$$z_{\frac{1}{4}} = 0{,}3176.$$

In Ausführung der Gleichung (31) erhält man schließlich die nachstehenden mittleren Fehler:

Zentralwert	$1{,}2534 \frac{\mu}{\sqrt{n}}$
1. und 9. Dezil	$1{,}7084 \frac{\mu}{\sqrt{n}}$
2. und 8. „	$1{,}4291 \frac{\mu}{\sqrt{n}}$
3. und 7. „	$1{,}3187 \frac{\mu}{\sqrt{n}}$
4. und 6. „	$1{,}2695 \frac{\mu}{\sqrt{n}}$
Quartil	$1{,}3634 \frac{\mu}{\sqrt{n}}$.

Beispiel. In dem in Art. 60 behandelten Fall der Körperhöhen amerikanischer Rekruten können die Voraussetzungen dieser Theorie als genügend erfüllt angesehen werden. Man hat da zur endgültigen Ausrechnung die Daten:

$$n = 25878 \qquad \mu = 2{,}5524 \text{ Zoll}$$

zu gebrauchen und erhält folgende Übersicht der Verteilung mit Angabe des mittleren Fehlers[1]):

	Zoll	Mittlerer Fehler (Zoll)
1. Dezil	63,374	0,0271
2. Dezil	64,454	0,0227
1. Quartil......................	64,882	0,0216
3. Dezil	65,270	0,0209
4. Dezil	66,013	0,0201
Zentralwert	66,651	0,0199
6. Dezil	67,323	0,0201
7. Dezil	68,041	0,0209
3. Quartil......................	68,454	0,0216
8. Dezil	68,867	0,0227
9. Dezil	70,066	0,0271

§ 4. Normale Korrelation.

131. Wenn man in den Feldmittelpunkten einer Korrelationstabelle, die sich auf zwei stetige Variable X_1, X_2 bezieht, zur Ebene der Tafel Lote errichtet und auf ihnen die Häufigkeiten der betreffenden Wertepaare nach einem geeigneten Maßstab aufträgt, so erhält man ein Punktsystem im Raume. Diesem passe man in derselben Weise, wie es bezüglich der Verteilung der Werte einer einzelnen Variablen mit der Häufigkeitskurve geschehen ist, eine krumme Fläche an. Ein Modell dieser Fläche würde eine anschauliche Vorstellung der Häufigkeitsverteilung der Wertpaare von X_1 und X_2, also auch von der zwischen ihnen etwa bestehenden Korrelation geben. Man wird daher der Fläche die Namen Häufigkeitsfläche und Korrelationsfläche je nach Sachlage beilegen.

Angenommen, die beiden Variablen zeigten, jede für sich betrachtet, normale Verteilung ihrer Werte, so daß der ersten die Verteilungsfunktion oder die Häufigkeitskurve

$$z_1 = z_{1,0}\, e^{-\frac{x_1^2}{2\mu_1^2}}$$

[1]) Dazu kommt der nicht näher bestimmbare Einfluß der Interpolation bei der Aufsuchung der Werte der ersten Kolonne, der aber gegenüber den Zahlen der zweiten Kolonne zurücktreten dürfte.

der zweiten die Kurve

$$z_2 = z_{2,0}\, e^{-\frac{x_2^2}{2\mu_2^2}}$$

entspricht; beide sollen bezogen sein auf den Punkt $(M_1 \mid M_2)$, der durch die arithmetischen Mittel von X_1 und X_2 bestimmt ist, so daß die x_1, x_2 die Abweichungen von diesen Mitteln und μ_1^2, μ_2^2 ihre mittleren Quadrate bedeuten.

Alsdann drückt sich die Häufigkeit der Wertverbindung $x_1 \mid x_2$, vorausgesetzt, daß die Variablen unabhängig, also auch korrelationslos sind, durch das Produkt der Häufigkeiten ihrer Komponenten aus, ist somit:

$$z = z_0\, e^{-\frac{1}{2}\left(\frac{x_1^2}{\mu_1^2} + \frac{x_2^2}{\mu_2^2}\right)} \tag{1}$$

Der Wert der Konstanten z_0 ist jedenfalls proportional dem Produkt $z_{1,0}\, z_{2,0}$ und der Proportionalitätsfaktor k ist so zu bestimmen, daß das Doppelintegral

$$\iint z\, d x_1 d x_2$$

über die ganze Ebene ausgedehnt den Wert N, d. i. der Umfang des Kollektivs, darstellt in derselben Weise, wie dies bei der Häufigkeitskurve der Fall ist; nun ist dieses Doppelintegral, von dem konstanten Faktor z_0 abgesehen, gleich dem Produkt der Integrale

$$\int_{-\infty}^{\infty} e^{-\frac{x_1^2}{2\mu_1^2}}\, d x_1 = \mu_1 \sqrt{2\pi} \qquad \int_{-\infty}^{\infty} e^{-\frac{x_2^2}{2\mu_2^2}}\, d x_2 = \mu_2 \sqrt{2\pi},$$

folglich bestimmt sich k aus dem Ansatze

$$k\, z_{1,0}\, z_{2,0} \cdot 2\pi \mu_1 \mu_2 = N,$$

und da $z_{1,0} = \frac{N}{\mu_1\sqrt{2\pi}}$, $z_{2,0} = \frac{N}{\mu_2\sqrt{2\pi}}$, so kommt $k = \frac{1}{N}$ und damit zugleich

$$\frac{N}{2\pi\mu_1\mu_2} \tag{2}$$

Über den Verlauf der Fläche sei folgendes bemerkt.

Den größten Wert verlangt z an der Stelle $x_1 = 0$, $x_2 = 0$, d. h. über dem Punkt $M_1 \mid M_2$ liegt der Gipfelpunkt der Fläche.

Ihre Schichtenlinien, d. s. die Projektionen ihrer Schnitte mit Ebenen parallel zur Grundebene, sind ähnliche und ähnlich liegende Ellipsen um $M_1 \mid M_2$ als gemeinsamen Mittelpunkt, enthalten in der Gleichung

$$\frac{x_1^2}{\mu_1^2} + \frac{x_2^2}{\mu_2^2} = u^2, \tag{3}$$

aus der das Achsenverhältnis $\mu_1 : \mu_2$ zu entnehmen ist.

Alle zur Grundebene senkrechten Schnitte sind Normalkurven. Von den Schnitten parallel zu den Achsen ist dies ohne weiteres zu erkennen; setzt man nämlich in (1) z. B. $x_2 = b$, so wird

$$z = z_0\, e^{-\frac{b^2}{2\mu_2^2}}\, e^{-\frac{x_1^2}{2\mu_1^2}}$$

und dies stellt — bei variabel gedachtem b — eine Schar von Normalkurven mit der gemeinsamen mittleren Abweichung μ_1 und variabler Gipfelordinate dar. Ähnlich für die Schnitte parallel zur x_2-Achse. Von einem Schnitt wie $x_2 = \alpha x_1 + \beta$ ist dies so nachzuweisen. Der absolute Betrag des Exponenten von e in (1) nimmt jetzt die Form an

$$\frac{x_1^2}{2\,\mu_1^2} + \frac{(\alpha x_1 + \beta)^2}{2\mu_2^2}$$

und läßt sich umgestalten in

$$\frac{(x_1 - A)^2 + B}{2\,C^2}$$

wo A, B, C aus μ_1, μ_2, α, β gebildete Ausdrücke bedeuten; mithin wird

$$z = z_0\, e^{-\frac{B}{2\,C^2}}\, e^{-\frac{(x_1 - A)^2}{2\,C^2}}$$

und dazu sind ähnliche Bemerkungen zu machen wie vorhin.

Übrigens kann man alle Flächen (1) auf eine einzige Grundform zurückführen, indem man die Abweichungen x_1 in der Einheit μ_1, die Abweichungen x_2 in der Einheit μ_2 ausdrückt und

$$\frac{x_1}{\mu_1} = \xi_1, \quad \frac{x_2}{\mu_2} = \xi_2 \tag{4}$$

setzt; damit wird

$$z = z_0\, e^{-\frac{\xi_1^2 + \xi_2^2}{2}} \tag{5}$$

zur Gleichung einer Rotationsfläche, deren Rotationsachse das Lot im Punkte $M_1 \mid M_2$ und deren Meridian die Kurve $z = z_0\, e^{-\frac{\xi^2}{2}}$ ist.

Die Substitution (4) bedeutet eine zweimalige affine Umformung der Fläche (1); auch daraus sind die erwähnten Schnitteigenschaften erkennbar.

132. Nun gehen wir von dem behandelten besonderen Fall zu dem allgemeinen über, daß X_1, X_2 in Korrelation stehen mit linearer Regression.

Die Regressions-Fehlergleichungen seien in der bereits vereinbarten Form

$$\begin{aligned} x_{1.2} &= x_1 - b_{12} x_2 \\ x_{2.1} &= x_2 - b_{21} x_1 \end{aligned} \tag{6}$$

geschrieben. Die Bestimmung von b_{12} geschieht aus der Bedingung

$$\Sigma\left(x_{1.2}^2\right) \text{ ein Minimum,}$$

ebenso die Bestimmung von b_{21} aus

$$\Sigma\left(x_{2.1}^2\right) \text{ ein Minimum.}$$

Die erste Bedingung, in Bezug auf b_{12} differenziert, gibt

$$\Sigma(x_{1.2} x_2) = 0,$$

die zweite, ebenso behandelt in Bezug auf b_{21}:

$$\Sigma(x_{2.1} x_1) = 0;$$

diese Normalgleichungen besagen, daß $x_{1.2}$ und x_2 und ebenso $x_{2.1}$ und x_1 korrelationslos sind.

Besteht zwischen diesen Paaren auch Unabhängigkeit und befolgen die einzelnen Größen das normale Verteilungsgesetz, so unterliegen ihre Wertverbindungen einem Gesetz von dem Bau (1), und zwar lautet dieses für das Paar $x_{1.2}$, x_2:

$$z = z_0 e^{-\frac{1}{2}\left(\frac{x_{1.2}^2}{\mu_{1.2}^2} + \frac{x_2^2}{\mu_2^2}\right)}$$

darin ist

$$z_0 = \frac{N}{2\pi \mu_{1.2} \mu_2}$$

Der Klammerausdruck im Exponenten von e entwickelt sich aber wie folgt:

$$\frac{(x_1 - b_{12} x_2)^2}{\mu_{1.2}^2} + \frac{x_2^2}{\mu_2^2} = \frac{x_1^2}{\mu_{1.2}^2} + \left(\frac{b_{12}^2}{\mu_{1.2}^2} + \frac{1}{\mu_2^2}\right) x_2^2 - \frac{2 b_{12} x_1 x_2}{\mu_{1.2}^2} =$$

$$= \frac{x_1^2}{\mu_{1.2}^2} + \frac{x_2^2}{\mu_{2.1}^2} - 2 r_{12} \frac{x_1 x_2}{\mu_{1.2} \mu_{2.1}},$$

wenn man in Rechnung bringt, daß

$$b_{12} = r_{12} \frac{\mu_1}{\mu_2} \qquad \mu_{1.2} = \mu_1 \left(1 - r_{12}^2\right)^{\frac{1}{2}} \qquad \mu_{2.1} = \mu_2 \left(1 - r_{12}^2\right)^{\frac{1}{2}}.$$

Zu derselben Schlußform des Klammerausdrucks kommt man, wenn man von dem Variablenpaar x_1, $x_{2.1}$ ausgeht.

Demnach wird schließlich [1])

$$z = z_0 e^{-\frac{1}{2}\left(\frac{x_1^2}{\mu_{1.2}^2} + \frac{x_2^2}{\mu_{2.1}^2} - 2r_{12}\frac{x_1 x_2}{\mu_{1.2}\mu_{2.1}}\right)} \tag{7}$$

mit

$$z_0 = \frac{N}{2\pi\mu_1\mu_2(1-r_{12}^2)^{\frac{1}{2}}}. \tag{8}$$

Die durch (7) dargestellte Korrelationsfläche ist aber von der früheren, im Falle der Unabhängigkeit gefundenen Fläche (1) nur der Lage nach verschieden; denn der Exponent von e in (7) läßt sich durch eine Drehungstransformation in die Form des Exponenten von (1) bringen.

Eine solche Transformation schreibt sich nämlich allgemein

$$x_1 = \xi_1 \cos\varphi - \xi_2 \sin\varphi$$
$$x_2 = \xi_2 \cos\varphi + \xi_1 \sin\varphi;$$

durch ihre Einführung verwandelt sich der eingeklammerte Faktor des Exponenten in (7) in ein Trinom mit ξ_1^2, ξ_2^2, $\xi_1\xi_2$, und zwar sind die Koeffizienten dieser Glieder in der gleichen Reihenfolge

$$\frac{\cos^2\varphi}{\mu_{1.2}^2} + \frac{\sin^2\varphi}{\mu_{2.1}^2} - 2r_{12}\frac{\sin\varphi\cos\varphi}{\mu_{1.2}\,\mu_{2.1}}$$

$$\frac{\sin^2\varphi}{\mu_{1.2}^2} + \frac{\cos^2\varphi}{\mu_{2.1}^2} + 2r_{12}\frac{\sin\varphi\cos\varphi}{\mu_{1.2}\,\mu_{2.1}}$$

$$-\frac{2\sin\varphi\cos\varphi}{\mu_{1.2}^2} + \frac{2\sin\varphi\cos\varphi}{\mu_{2.1}^2} - 2r_{12}\frac{\cos^2\varphi - \sin^2\varphi}{\mu_{1.2}\,\mu_{2.1}};$$

verschwindet der dritte Koeffizient, so ist die Form erzielt; der dazu nötige Drehungswinkel bestimmt sich also aus der Gleichung

$$\left(\mu_1^2 - \mu_2^2\right)\sin 2\varphi - 2r_{12}\mu_1\mu_2\cos 2\varphi = 0, \tag{9}$$

aus der

$$\operatorname{tg} 2\varphi = \frac{2r_{12}\mu_1\mu_2}{\mu_1^2 - \mu_2^2} \tag{10}$$

folgt.

---

[1]) Zur Ableitung dieser Formel und der entsprechenden für beliebig viele Variable sei auf die Abhandlung von A. Guldberg: On the Law of Errors in the Space of p Dimensions, The Tôhoku Mathematical Journal, Bd. 17, 1920, p. 18—23 hingewiesen. — Vgl. ferner A. Guldberg, On the Theory of Frequency-Distributions, Skandinavsk Aktuarietidskrift, Jahrg. 1919, p. 224—232. Diese letztere Arbeit betrifft die Zurückführung beliebiger Verteilungen einer Variablen auf die Normalkurve, ebenso beliebiger Verteilungen zweier Variablen auf die Normalfläche und die Ausdehnung dieser Analyse auf beliebig viele Variable.

Der Exponent erlangt dann die Form

$$-\frac{1}{2}\left(\frac{\xi_1^2}{M_1^2}+\frac{\xi_2^2}{M_2^2}\right)$$

und zwar ist

$$\frac{1}{M_1^2}=\frac{\cos^2\varphi}{\mu_{1.2}^2}+\frac{\sin^2\varphi}{\mu_{2.1}^2}-2r_{12}\frac{\sin\varphi\cos\varphi}{\mu_{1.2}\,\mu_{2.1}}$$

$$\frac{1}{M_2^2}=\frac{\sin^2\varphi}{\mu_{1.2}^2}+\frac{\cos^2\varphi}{\mu_{2.1}^2}+2r_{12}\frac{\sin\varphi\cos\varphi}{\mu_{1.2}\,\mu_{2.1}};$$

daraus ergibt sich durch Addition

$$\frac{1}{M_1^2}+\frac{1}{M_2^2}=\frac{M_1^2+M_2^2}{M_1^2\,M_2^2}=\frac{1}{\mu_{1.2}^2}+\frac{1}{\mu_{2.1}^2}=\frac{\mu_1^2+\mu_2^2}{\mu_1^2\,\mu_2^2\left(1-r_{12}^2\right)},$$

was dadurch erfüllt werden kann, daß man setzt

$$M_1^2+M_2^2=\mu_1^2+\mu_2^2 \tag{11}$$

$$M_1^2\,M_2^2=\mu_1^2\,\mu_2^2\left(1-r_{12}^2\right); \tag{12}$$

ferner durch Subtraktion

$$\frac{1}{M_2^2}-\frac{1}{M_1^2}=\frac{M_1^2-M_2^2}{M_1^2\,M_2^2}=\frac{\left(\mu_1^2-\mu_2^2\right)\cos 2\varphi+2r_{12}\mu_1\mu_2\sin 2\varphi}{\mu_1^2\,\mu_2^2\left(1-r_{12}^2\right)},$$

was sich, wenn man im Zähler r_{12} durch den aus (9) resultierenden Ausdruck ersetzt, weiter verwandelt in

$$\frac{M_1^2-M_2^2}{M_1^2\,M_2^2}=\frac{\mu_1^2-\mu_2^2}{\mu_1^2\,\mu_2^2\left(1-r_{12}^2\right)\cos 2\varphi};$$

mit Rücksicht auf (12) folgt aber daraus

$$M_1^2-M_2^2=\frac{\mu_1^2-\mu_2^2}{\cos 2\varphi} \tag{13}$$

und mit Zuziehung von (10)

$$\sin 2\varphi=\frac{2\,r_{12}\,\mu_1\,\mu_2}{M_1^2-M_2^2}. \tag{14}$$

Da 2φ auf das Intervall $(0,\pi)$ beschränkt bleiben kann, worin $\sin 2\varphi>0$ ist, so sieht man, daß das Vorzeichen von $M_1^2-M_2^2$ mit dem Vorzeichen von r_{12} übereinstimmt. Dadurch ist die Lösung des Gleichungspaares (11), (13) eindeutig bestimmt. Es fallen nämlich bei positiver Korrelation die großen Achsen der Ellipsen in die Richtung ξ_1, bei negativer Korrelation in die Richtung ξ_2.

Hat die vorstehende Transformation und im besonderen die Gleichung (10) zu den Hauptachsen der Ellipsen und damit auch zu den Hauptebenen der Korrelationsfläche geführt, in Bezug auf welche sie orthogonal symmetrisch ist, so handelt es

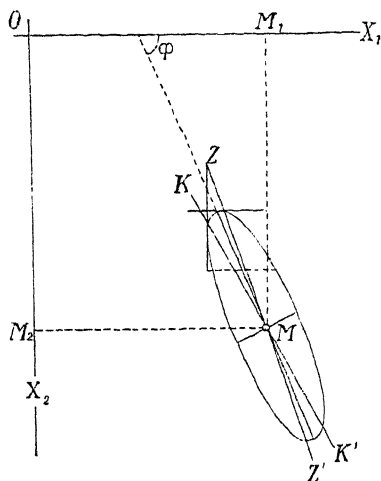



Fig. 38. Normale Korrelation.

sich noch darum, in welcher Beziehung die Regressionsgeraden zu den Ellipsen stehen. Diese Linien sind die Orte der arithmetischen Mittel der Kolonnen und der Zeilen; da nun alle vertikalen Schnitte der Fläche Normalkurven sind, so fallen die genannten arithmetischen Mittel unter die Gipfelpunkte der Schnittkurven, also in die Mitten der Ellipsensehnen, welche der x_2-, bzw. der x_1-Achse parallel sind, d. h. die Gerade KK' der Kolonnenmittel ist der zur X_2-Achse konjugierte Durchmesser des Ellipsensystems, und die Gerade ZZ' der zur X_1-Achse konjugierte Durchmesser, Fig. 38.

Die Projektionen der Schichtenlinien der Korrelationsfläche haben für die Korrelationstafel die Bedeutung von Linien gleicher Häufigkeit der auf ihnen liegenden Wertepaare X_1/X_2.

Wie nun die Schichtenlinien einer Terrainfläche, wenn die zugehörigen Ebenen in gleichen Abständen gelegt sind, in ihrer mehr oder weniger dichten Lagerung die größere oder kleinere Steilheit der Fläche zeigen, so ist bei der Korrelationsfläche unter gleichen Umständen die größere oder geringere Häufung aus der Lagerung der Linien gleicher Häufigkeit zu entnehmen.

Ersetzt man in der transformierten Flächengleichung

$$z = z_0 \, e^{-\frac{1}{2}\left(\frac{\xi_1^2}{M_1^2} + \frac{\xi_2^2}{M_2^2}\right)}$$

den Klammerausdruck durch u^2 und erteilt z einen bestimmten, der Tafel nach zulässigen Wert, so ergibt sich aus der Gleichung

$$z = z_0 \, e^{-\frac{u^2}{2}}$$

der diesem z zugehörige Wert von u^2 und damit die Gleichung

$$\frac{\xi_1^2}{M_1^2} + \frac{\xi_2^2}{M_2^2} = u^2;$$

erteilt man z eine Reihe äquidistanter Werte, so führen die entsprechenden u^2-Werte zu den diesen Häufigkeiten korrespondierenden Ellipsen.[1])

[1]) Am Schlusse des Art. 76 ist auf Wirths Auffassung der linearen Korrelation hingewiesen worden. Dieser zufolge ist es die erste Hauptachse des Systems der Ellipsen gleicher Häufigkeit (Isoplethieen nennt er sie), welche die Rolle seiner „mittleren Geraden“ spielt. In der Tat führt das Problem, die Gerade $-X\sin\varphi + Y\cos\varphi - p = 0$ so zu bestimmen, daß die Bildpunkte des zweidimensionalen Kollektivs Abstände von ihr haben, deren Quadratsumme ein Extrem ist, auf die Hauptachsen des Ellipsensystems, insbesondere gehört zum Minimum die erste (die große) Hauptachse. Gerade in dieser Beziehung zur normalen Korrelation tritt die bevorzugte Stellung, welche Wirths „mittlere Gerade“ gegenüber den Regressionslinien einnimmt, deutlich hervor, und dieser Vorzug ist ihr auch unabhängig davon, also allgemein, einzuräumen.

133. Beispiel. Es soll geprüft werden, ob die Korrelation zwischen der Länge und Breite der Blätter des wilden Efeu (Art. 72, Tab. 54; Art. 79, 5) normalen Charakter hat.

Die dort gefundenen Daten sind:

$$N = 2500, \quad M_1 = 10{,}90, \quad M_2 = 13{,}23, \quad \mu_1 = 1{,}71 \text{ Klassen} = 3{,}42 \text{ Achtelzoll}$$
$$\mu_2 = 2{,}28 \quad „ \quad = 4{,}56 \quad „$$
$$r_{12} = 0{,}86.$$

Da Formel (10) für tg 2φ einen negativen Wert gibt, so ist 2φ ein stumpfer Winkel, u. zw. $108^0\,44'$, folglich

$$\varphi = 54^0\,22'.$$

Aus (14) schließt man, daß $\mathsf{M}_1 > \mathsf{M}_2$; die gültige Lösung von

$$\mathsf{M}_1^2 + \mathsf{M}_2^2 = 32{,}49$$
$$\mathsf{M}_1^2\,\mathsf{M}_2^2 = 63{,}33$$

ist also $\mathsf{M}_1 = 5{,}51$, $\mathsf{M}_2 = 1{,}45$ Achtelzoll.

Bei der Berechnung von z_0 nach Formel (8) sind μ_1, μ_2 in Klassengröße zu verwenden, weil sich die Häufigkeiten auf Klassenintervalle beziehen; man findet

$$z_0 = 200{,}0.$$

Demnach lautet die Gleichung der Korrelationsfläche in transformierter Gestalt

$$z = 200{,}0\, e^{-\frac{1}{2}\left(\frac{\xi_1^2}{(5{,}51)^2} + \frac{\xi_2^2}{(1{,}45)^2}\right)}; \tag{a}$$

für die weitere Prüfung ist jedoch die ursprüngliche Form (7) zweckmäßiger; zu ihrer Herstellung braucht man $\mu_{1.2}$ und $\mu_{2.1}$ und findet dafür

$$\mu_{1.2} = 1{,}75 \qquad \mu_{2.1} = 2{,}33 \text{ Achtelzoll},$$

die Gleichung selbst also lautet

$$z = 200{,}0\, e^{-\frac{1}{2}\left(\frac{x_1^2}{(1{,}75)^2} + \frac{x_2^2}{(2{,}33)^2} - \frac{x_1\,x_2}{2{,}36}\right)}. \tag{b}$$

Die Prüfung des Anschlusses der Fläche an das Beobachtungsmaterial kann in verschiedener Art erfolgen.

1) Man konstruiert mit Hilfe des Punktes M_1/M_2 und des Winkels φ das Hauptachsenkreuz und zeichnet auf Grund von (a) einige Ellipsen gleicher, im voraus gewählter Häufigkeit und sieht nach, ob diese Ellipsen die Zeilen und Kolonnen an Stellen durchsetzen, welchen die betreffende Häufigkeit zukommt, was im allgemeinen eine Interpolation erfordert. Man kann auch in den Zeilen und Kolonnen die Punkte feststellen, zu welchen die angenommene Häufigkeit gehört, sie zu einem Polygon verbinden und zusehen, wie sich das Polygon zu der Ellipse verhält, von der es eine Näherung sein soll.

2) Man rechnet mit Hilfe von (b) die Häufigkeiten für die einzelnen Felder und vergleicht sie mit den wirklich beobachteten.

Zu dem einen wie dem andern gleich mühsamen Verfahren wird man nur greifen, wenn es sich um eine Tafel großen Umfangs handelt, weil nur unter dieser Voraussetzung ein befriedigender Anschluß zu erwarten ist.

Die folgende Probe zeigt die berechneten Häufigkeiten einiger zentralen Felder unserer Tafel und darunter in Klammern die beobachteten. Trotz erheblicher Abweichungen im einzelnen kann von einem gleichartigen Verlauf der beiden Zahlengruppen gesprochen werden.

Breite \ Länge	5,95	7,95	9,95	11,95	13,95
7,95		100 (106)	38 (37)	4 (4)	
9,95	42 (33)	139 (190)	124 (152)	30 (31)	2 (4)
11,95	12 (7)	92 (88)	192 (227)	108 (98)	16 (16)
13,95	2 (1)	29 (26)	142 (137)	187 (216)	66 (66)
15,95		4 (.)	51 (55)	155 (136)	

Die Summe der gerechneten Häufigkeiten innerhalb dieses Bereiches ist 1535, die der beobachteten 1630.

Besonders an manchen anthropologischen Materien ist die Tendenz zur normalen Korrelation beobachtet worden. Wir heben nachstehend das Kernstück aus einer Tabelle von K. Pearson und A. Lee [1]), betreffend die Korrelation zwischen der Körpergröße des Vaters und des Sohnes heraus und geben die Elemente der Rechnung an.

[1]) K. Pearson und A. Lee, „On the Laws of Inheritance in Man". Biometrika, vol. II, 1903, p. 415. Vgl. G. U. Yule, An Introduction to the Theory of Statistics. London 1932, p. 160 und 326. Die Korrelation zwischen der Körpergröße des Vaters und der des Sohnes hat neuerdings M. Boldrini einer Untersuchung unterzogen. Vgl. M. Boldrini, La Fertilità dei Biotipi. Mailand 1931. Public. della Università Catholica del Sacro Cuore. Vol. IV; vgl. hierzu E. Weber, Einführung in die Variations- und Erblichkeits-Statistik. München 1935, S. 229 und F. Ringleb, Mathematische Methoden der Biologie, Leipzig 1937, S. 178.

$$N = 1078, \quad M_1 = 67{,}70 \quad \mu_1 = 2{,}72 \text{ Zoll (Vater)}$$
$$M_2 = 68{,}66 \quad \mu_2 = 2{,}75 \quad \text{„ (Sohn)}$$
$$r_{12} = 0{,}51;$$

daraus geht die Gleichung

$$z = 26{,}7\, e^{-\frac{1}{2}\left(\frac{x_1^2}{5{,}47} + \frac{x_2^2}{5{,}60} - \frac{x_1 x_2}{5{,}43}\right)}$$

hervor und Winkel φ beträgt $45^0\ 37'$. Die berechneten und beobachteten Häufigkeiten gestalten sich wie folgt:

Sohn \ Vater	66″	67″	68″	69″	70″	
67″		22,2 (25,75)	19,8 (19,5)	14,7 (12,5)		56,7 (57,75)
68″	21,9 (24,25)	25,6 (31,5)	25,0 (23,5)	20,3 (29,5)	13,8 (13,25)	106,6 (122,0)
69″	19,2 (18,25)	24,7 (16)	26,5 (24)	23,6 (29)	17,5 (21,5)	111,5 (108,75)
70″	14,2 (18,75)	20,0 (11,75)	23,4 (19,5)	22,9 (22,5)	18,6 (19,5)	99,1 (92,0)
71″		13,5 (10,75)	17,3 (19)	18,6 (14,75)		49,4 (44,5)
	55,3 (61,25)	106,0 (95,75)	112,0 (105,5)	100,1 (108,25)	49,9 (54,25)	423,3 (425)

Der gute Anschluß zeigt sich besonders in den Summen der Häufigkeitszahlen: die gerechneten ergeben zusammen 423,3, die beobachteten 425.

SACHREGISTER.

(Die Zahlen beziehen sich auf Seiten.)

NAMENREGISTER.